National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file *Votre référence*

Our file *Notre référence*

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

MOLECULAR CHARACTERIZATION OF A SINGLE *cab* GENE IN LODGEPOLE PINE

BY

JOHN BARRETT

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

DEPARTMENT OF FOREST SCIENCE

Edmonton, Alberta
Fall 1992

National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-77244-1

Canada

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: John Barrett

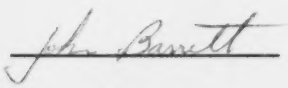
TITLE OF THESIS: Molecular Characterization of a Chlorophyll a/b (*cab*) Gene in Lodgepole Pine

DEGREE: Doctor of Philosophy

YEAR THIS DEGREE GRANTED: Fall 1992

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



366 Elizabeth St.
Sault Ste. Marie, ON
P6B 3H4

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Molecular Characterization of a Single *cab* Gene in Lodgepole Pine submitted by John Barrett in parital fulfillment of the requirements for the degree of Doctor of Philosophy.

Bruce Dancik, Co-supervisor

Curtis Strobeck, Co-supervisor

Ross Hodgetts, Committee Member

Allen Good, Committee Member

for Ron Sederoff, External Examiner

ABSTRACT

A single copy of a Type I chlorophyll a/b binding protein (*cab)* gene from lodgepole pine (*Pinus contorta* Dougl., var. *latifolia* (Engelm.) has been isolated and sequenced. The gene encodes a proposed 40 amino acid transit peptide and a 234 amino acid mature CAB polypeptide. There is no evidence of any introns and examination of the Type-specific amino acids confirms that this gene is a Type I *cab* copy. The coding region is highly conserved compared to a similar gene from scots pine, the strong homology extending into the 5' and 3' untranslated regions. Sequence comparison of the 1 kb region upstream of the translation start codon from the lodgepole pine *cab* gene exhibited little similarity to identified regulatory regions in pea, tobacco, wheat and *Arabidopsis cab* genes. The *cab* gene, which is 62% G/C rich, is surrounded by 70-75% A/T rich regions which may confer some regulatory feature on its expression.

Comparison of the transit peptides of all other Photosystem II (PSII) *cab* genes suggests that they are divided into four classes. These would include the established Type I and Type II transit peptides, the woody plant Type I transit peptides and the transit peptides of the lower plants. Analysis of the amino acid composition across these groups indicates two small conserved motifs. The first, which is shared loosely by 90% of the PSII *cab* genes, is F/L-A-G and may be important for chloroplast membrane receptor recognition. The second region of similarity lies at the proposed cleavage site between the transit peptide and the mature polypeptide. The specific structural properties are maintained in 93% of the *cab* PSII transit peptides. This may be an important recognition site for the stromal peptidase.

Phylogenetic analysis of the PSII *cab* genes suggests that the evolution of the Type I and Type II copies occurred before the divergence of the angiosperms from the gymnosperms. As well, moss and fern contain genetic structures that suggest that they may be descendants of an ancestral copy which contains features of both the Type I and Type II *cab* genes.

ACKNOWLEDGEMENTS

I would like to thank my two supervisors, Dr. Bruce Dancik of the Dept. of Forest Science who, along with Forestry Canada, NSERC, the Olds Agricultural Society and the Dept. of Forest Science, provided financial support, and to Curtis Strobeck of the Dept. of Zoology who supplied lab support.

Thanks to the members of the Strobeck laboratory, in particular, Dr. Robin Beech for good friendship and good science twenty-four hours a day.

Within the Dept. of Genetics I would like to acknowledge the technical assistance of Bill Clark and to extend great thanks to Dr. Ross Hodgetts, who was there from the beginning and to Dr. Allan Good who came in at the end. Both contributed strongly to the final product.

Thanks to Dr. Basil Arif and the members of his laboratory at the Forest Pest Management Institute in Sault Ste. Marie, Ontario for their encouragement and support.

I wish to acknowledge and thank my family and my in-laws for their strong encouragement and support.

Finally to Carol (and Michael, Kimberly and Karen) you exhibited great patience and an unbelieveable ability to keep this family going with so little. Thanks for the love support, encouragement and patience.

TABLE OF CONTENTS

CHAPTER ONE 1
GENERAL INTRODUCTION 1
CAB Protein Transport 4
Transmembrane Structure 7
Thylakoid Insertion 8
Conservation And Evolution Of The cab *Gene Sequences* 8
OBJECTIVES OF DISSERTATION 9
LITERATURE CITED 12

CHAPTER TWO 14
INTRODUCTION 14
MATERIALS AND METHODS 16
DNA Extraction and Restriction Digestion 16
Hybridization conditions 17
Library Construction 18
Preparation of Lambda Arms 18
Preparation of Packaging Extracts 19
In vitro *Packaging* 19
Bacterial Cultures 20
Preparation of Cloned DNA 21
Construction of Deletions 23
Sequencing Single-stranded DNA 24
Sequencing Double-stranded DNA 24
Sequencing Gels 25
RESULTS 25
λPNcAB10-1 and pTZ19R-5 26
DESCRIPTION OF THE OTHER λPNcAB CLONES 30
λPNcAB4-10 30
λPNcAB5-10 35
λPNcAB7-10 37
λPNcAB7-20 37
DISCUSSION 47
LITERATURE CITED 51

CHAPTER THREE 54
INTRODUCTION 54
MATERIALS AND METHODS 56
Construction of Packaging Extract and Lambda Arms 56
Extraction of Lodgepole Pine Genomic DNA 57
Ligation Reaction 57
Library Construction 58
Hybridization and Identification of Positives 58
Growth of Phage and DNA Extraction 59

Restriction Mapping 59
Sequencing 60
RESULTS 61
cab *Coding Region* 61
Transit Peptide 66
Untranslated Regions 67
CpG Richness 71
CpG Dinucleotides 73
Promoter Region 73
DISCUSSION 74
LITERATURE CITED 77

CHAPTER FOUR 83
INTRODUCTION 83
MATERIALS AND METHODS 85
Sequences 85
Hydrophilicity Analysis 85
Hydrophobic Moment Analysis 86
RESULTS 86
Lodgepole Pine Transit Peptide 86
Amino Acid Frequency Analysis 88
Hydrophobic Moment 91
DISCUSSION 96
CAB Transit Peptides 96
Type I CAB Transit Peptides 96
Type II CAB Transit Peptides 97
Woody Plant Transit peptides 97
Transit Peptides of Lower Plants 98
Comparison of The Transit Peptides Between the Groups 99
Mechanisms of Transport 101
Model of Transportation For LHC 102
LITERATURE CITED 103

CHAPTER FIVE 107
INTRODUCTION 107
MATERIALS AND METHODS 108
PAUP Analysis 108
Distance Matrix 110
Model Assumptions 110
RESULTS 111
DNA Parsimony 111
Distance Analysis 118
DISCUSSION 120
LITERATURE CITED 126

CHAPTER SIX 130

Future Research 133
LITERATURE CITED 135

LIST OF TABLES

Table 1.1. Organization of the Chlorophyll a/b Binding Protein and Their Respective Genes in Tomato..... 3

Table 1.2. Diagnostic Amino Acids Separating Type I and Type II CAB Proteins..... 10

Table 3.1. Comparison of the 705 bp cab Coding Region for the Lodgepole Pine Clone λPNcAB2 Against Selected Other *cab* Genes..... 65

Table 4.1. Compilation of CAB Transit Peptides..... 87

Table 5.1. Taxonomic Classification of the Plants Used in This Study..... 112

LIST OF FIGURES

Figure 1.1. Evolutionary relationship of the chorophyll a/b binding proteins and their genes from tomato..... 5

Figure 1.2 Schematic illustrating intron positions in generalized models of the various *cab* genes from tomato..... 6

Figure 2.1 Restriction map of the lambda clone PNcAB10-1..... 27

Figure 2.2. Restriction map of pTZ19R-5..... 29

Figure 2.3. Identification of positive λPNcAB clones probed with homologous probe..... 31

Figure 2.4. Southern analysis of λPNcAB clones probed with pCab3..... 32

Figure 2.5. Map of lodgepole pine Type I *cab* gene, PNcAB2..... 33

Figure 2.6. Restriction map of lambda clone 4-10..... 34

Figure 2.7. Restriction map of lambda clone 5-10..... 36

Figure 2.8. Restriction map of lambda clone 7-10..... 38

Figure 2.9. Comparison of the sequence of λPNcAB7-10 and the *cab* gene PNcAB2 from lodgepole pine..... 39

Figure 2.10. Restriction map of λPNcAB7-20..... 40

Figure 2.11. Comparison of the sequence from clone 7-20 which a 44 nucleotide region with other *cab* genes..... 42

Figure 2.12. Lodgepole pine clones probed with 282 bp *cab* gene..... 44

Figure 2.13. Autoradiogram of lodgepole pine clones probed with pTZ19R-5..... 45

Figure 3.1. The organization and sequencing strategy of the *cab* gene in lodgepole pine..... 62

Figure 3.2. The nucleotide and amino acid sequence of the lodgepole pine *cab* gene..... 63

Figure 3.3. Comparison of possible initiation start sites in six *cab* genes

and one ferredoxin gene..... 69

Figure 3.4. Sequence comparison of the 5' and 3' untranslated flanking regions of *cab* Type I genes..... 70

Figure 3.5. A/T composition profile of the lodgepole pine *cab* gene..... 72

Figure 3.6. Comparison of the promoter region between the CAAT and TATA box for selected *cab* genes..... 75

Figure 4.1. Hydrophilicity profile of lodgepole pine transit peptide and surrounding region..... 89

Figure 4.2a. Amphiphilic profiles of the lodgepole pine transit peptide region.. 92

Figure 4.2b. Amphiphilic profiles of the transit peptide of Petunia 22R..... 93

Figure 4.2c. Amphiphilic profiles of the transit peptide region in rice..... 94

Figure 4.2d. Amphiphilic profiles of the apple Type I *cab* gene..... 95

Figure 5.1. Cladogram representing the phylogenetic relationships of the PSII *cab* genes..... 114

Figure 5.2. Phylogenetic relationships of the PSII *cab* genes using 101 informative sites..... 116

Figure 5.3. Phylogenetic relationships of the PSII *cab* genes including the algae species..... 117

Figure 5.4. Phenetic analysis of 20 selected *cab* genes..... 119

Figure 5.5. Phylogenetic tree derived from KITSCH analysis of PSII *cab* genes..... 121

LIST OF ABBREVIATIONS

PSI, PSII; Photosystem I and II

LHC-I, LHC-II; Light harvesting complex I and II

SSU; small subunit of ribulosebisphosphate carboxylase

ug; microgram

mg; milligram

kb; kilobase

nt; nucleotide

ul; microlitre

U; unit

PNcAB; *Pinus contorta* chlorophyll a/b-binding gene

MW; molecular weight

CHAPTER ONE

GENERAL INTRODUCTION

Photosynthesis involves the conversion of light energy into the chemical energy used to transfer electrons from H_2O to $NADP^+$ and to form the phosphoanhydride bonds of ATP. The efficient capture and transfer of this light energy is mediated by the light absorbing photosystems. The chloroplasts of higher plants have two photosystems (PSI and PSII), both of which undergo charge separation but which differ in function. These photosystems are located within the thylakoid membranes and each photosystem is composed of chlorophyll molecules, carotenoids, cytochromes and other electron-transporting proteins. PSII is located in the stacked membranes of the grana while PSI is located in the non-stacked regions. Within each photosystem only two specialized reaction-centre chlorophyll *a* molecules, termed P680 in PSII and P700 in PSI, are capable of undergoing light driven electron transfer. The light-harvesting complex (LHC) is part of each photosystem. The LHC has no catalytic activity; its primary function is to capture light energy and transfer it to PSI or PSII. Each LHC contains three transmembrane polypeptides of 25,000-27,000 MW (Kuhlbrandt and Wang, 1991) which bind up to one-half of the chlorophyll *a* molecules, and nearly all of the chlorophyll *b* molecules. Most LHC are associated with PSII, nevertheless PSI has a specific LHC. The absorbed energy is then transferred rapidly ($<10^{-9}$ s) to the specialized chlorophyll

a molecules in the reaction centres of each photosystem. Within this framework, the role of the LHC may be to bind the chlorophyll molecules together and maintain the orientation of the chlorophyll molecules so that the process may occur efficiently. A recent model of the pea LHCII, based on electron crystallography, verifies the three-helix transmembrane model for polypeptides inserted into the thylakoids (Karlin-Neumann et al., 1985) and finds that the 15 chlorophyll molecules are organized in two levels, with rings roughly perpendicular to the membrane plane (Kuhlbrandt and Wang, 1991).

The light harvesting complexes (LHC) are composed of a very diverse family of proteins that are encoded by an equally diverse gene family. The genes encoding the peptides which form the light harvesting complexes of PSI and PSII are known as the *cab* genes (for chlorophyll a/b binding) and are well studied in a wide and increasing range of green plant species. The genes involved in the interception and transfer of light energy have been organized into groups based on similarity of gene structure and function of the gene product. For, example, within the two photosystems it has been demonstrated that there are four separate protein complexes involved in the interception of light energy. The polypeptides of CP24 and CP29 are associated with the photosynthetic core molecules within PSII. The chlorophyll a/b binding proteins are involved in energy transfer and the numerous polypeptides are separated into PSI and PSII (Table 1.1). Little is known about the structure of the CP24 and CP29 genes, however sequencing has confirmed that they share some common amino acid sequences.

Table 1.1. Organization of the chlorophyll a/b binding proteins and their respective genes in tomato. Adapted from Green et al. 1991.

Complex	Location	Chl a/b ratio	Polypeptide	Gene Types	No. gene copies	No. introns
LHCII	Major antenna	1.2	2 major	Type I	8	0
	PSII			Type II	2	1
			1 minor	Type III	ND	-
CP29	Core antenna	4-5	2	Type I	1	5
	PSII			Type II	ND	-
CP24	Minor PSII	<1	1-2	-	2	1
LHCI	PSI antenna	3-5	4	Type I	2	3
				Type II	1	4
				Type III	1	2
				Type IV	2	2

ND= not determined

The best characterized genes are those coding for the chlorophyll a/b binding proteins. These genes are organized on the basis of their location within a particular photosystem, the occurrence and location of any introns, and by sequence similarity. It has been well established that all these polypeptides share some structural homology and the shared features may be incorporated into an evolutionary tree (Fig. 1.1).

With the exception of the LHCII Type I cab genes, all other characterized light harvesting genes from tomato contain at least one intron (Fig. 1.2) and many of these gene types are composed of more than one gene copy (Green et al., 1991). Although the LHCII Type I and II genes are well characterized, information about the other light harvesting genes is based upon one or two nucleotide sequences. Nevertheless, of the ten intron positions illustrated in Fig. 1.2, only three are common in more than one gene type. The scattered appearance of the introns, as well as their occurrence within conserved regions suggest that the introns appeared after the divergence of the gene types (Green et al. 1991).

CAB Protein Transport

Cab genes constitute one of those unique gene families which resides in the nuclear genome, is cytoplasmically translated and processed, and whose product is then transported into the chloroplast. Transport of the mature CAB protein across the chloroplast envelope is facilitated by a small transit peptide, whose genetic information is encoded at the N-terminal portion of the *cab* gene. The actual mechanism of transport of this apoprotein is not known completely. Nevertheless, it appears to involve an

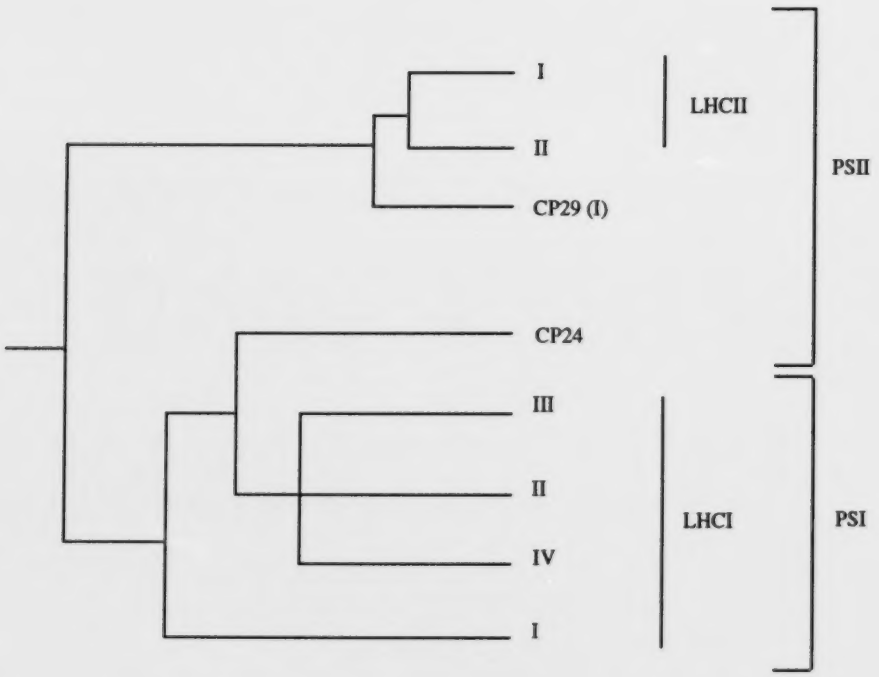


Figure 1.1. Evolutionary relationship of the chlorophyll a/b binding proteins and their genes from tomato. Adapted from Green et al. 1991.

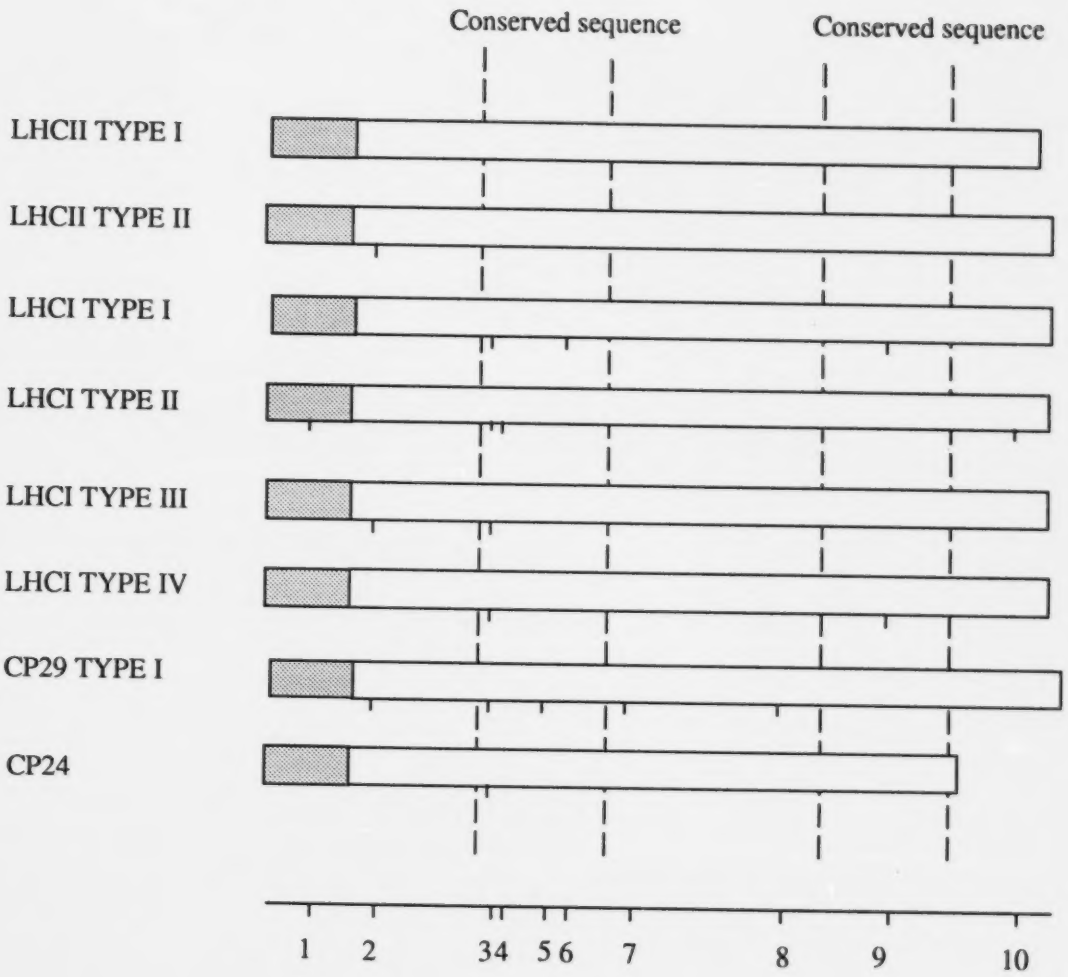


Figure 1.2. Schematic illustrating intron positions in generalized models of the various *cab* genes from tomato. Note that only three introns (2,3,9) appear more than once. Hatched areas indicate transit peptides and the relative sizes of the genes are represented by the boxes (modified from Green et al., 1991).

interaction between a chloroplast membrane surface receptor (Pain et al., 1988) and the transit peptide. Following transport the transit peptide is cleaved by a stromal peptidase. The mature polypeptide is then routed to the correct internal position based on some other recognition signal or mechanism. It might be expected that if the primary structure of the transit peptide was receptor specific the same genes might utilize the same transit peptide amino acid sequence. However, there is a lack of sequence similarity between the transit peptides for different types of *cab* genes within the same species. The general scheme suggests that the amino acid sequence of the transit peptides is similar for the same *cab* gene type in different plants. But they diverge significantly for different gene types within the same species.

Transmembrane Structure

It has been proposed that the mature CAB protein consists of three transmembrane domains which anchor the monomer within the thylakoid membrane (Karlin-Neumann et al., 1985). Hydrophobic analysis of the deduced amino acid sequence supports this model. Recently, it has been proposed that the order of the three helices in the polypeptide seems to be B, C and A, beginning at the N-terminus (Kuhlbrandt and Wang, 1991). Based on electron crystallography, it has been proposed that helices A and B are longer than C, are maintained at a 30° angle within the lipid bilayer and may extend into the stroma. Helix C is shorter, confined to the membrane and lies perpendicular to the membrane proper. The primary role of helices A and B seems to act as scaffolding for

packing the chlorophyll molecules into a small volume, ensuring optimal spacing and orientation for energy transfer.

Thylakoid Insertion

The mechanism and recognition of the thylakoid insertion site has not been elucidated. Nevertheless, the insertion step is probably directed by an unidentified topogenic sequence found in the mature protein (Keegstra, 1989). Although the exact location has not been identified it may be associated with the third membrane spanning helix (helix A, using Kuhlbrandt and Wang's notation, 1991; Kohorn et al., 1986).

Conservation And Evolution Of The cab *Gene Sequences*

Based on studies in tomato there appear to be eight different types of *cab* genes (Green et al., 1991). The Type I and II genes of LHCII are the most common and based on sequence comparison the most closely related. The nucleotide sequences encoding the mature protein are approximately 85% similiar, although the transit peptides are only 50% identical. It has been observed that the greatest variation occurs at the 5' end of the gene, encompassing the first 60-65 nucleotides.

The CAB proteins of the other light harvesting genes are substantially diverged (60-70%) suggesting that the gene duplications that gave rise to them occurred very early in the evolution of *cab* containing organisms. In contrast, the deduced amino acid sequence of the LHCII Type I and II CAB proteins is very well conserved indicating a much later gene duplication event. It has been noted (Demmin et al., 1989; Jansson and

Gustafsson 1990) that there are 15 diagnostic amino acids positions within the mature peptide which may be used to distinguish the two types (Table 1.2).

OBJECTIVES OF DISSERTATION

Preliminary evidence on the organization and structure of genes in gymnosperms appears to differ from angiosperms. For example, it has been demonstrated that conifers have very large nuclear genomes with a high proportion of repetitive DNA (Rake et al. 1980). Also, ribosomal DNA repeats in white spruce have been shown to be 18 kb (Beech, personal communication) and 35-43 kb in black spruce and red spruce (Bobola et al. 1992). This is in contrast to the 8-10 kb repeats in most angiosperms. As well, chloroplast organllar DNA is transmitted paternally in Douglas-fir (Neale et al. 1986) in contrast to the maternal inheritance in angiosperms. Conifers have also been shown to contain only a single region of rDNA on the chloroplast chromosome (Strauss et al. 1988) versus the repeated regions found in angiosperms chloroplasts.

Dramatic physiological differences have also been found in conifers. For example, contrary to angiosperms, conifers have been shown to synthesize chlorophyll in the dark and initial studies suggest that rubisco and cab genes are not strongly light regulated.

Clearly it is not sufficient to use the angiosperm model to explain conifer gene organization and structure. If serious work on the manipulation of gene products involved in photosynthesis and secondary metabolism are to successful it will be necessary to determine more about important conifers genes. As a first step toward establishing

Table 1.2. Diagnostic amino acids separating the Type I and Type II CAB proteins in photosystem II. Amino acid position refers to a consensus Type I mature LHCII protein. Amino acids are presented in their single letter codes. Adapted from Demmin et al. (1989) and Jansson and Gustafsson (1990).

Amino Acid Position	Type I	Type II
3	K	R
14	G,S	E,Q
16	P	I
23	V	P
31	G	E
32	E	Q
33	S,P,A	T
88	R	K
118	S	N
145	A	G
152	V,I,A	G
153	V,I,T,E	L
161	S	A
208	L	I
219	N	A

this information it is necessary to select a small group of important genes, whose function is well understood, and determine their structure and organization in a model conifer species. I am interested in the organization of the *cab* genes in lodgepole pine (*Pinus contorta* var. *latifolia* (Engelm.)). These genes have been shown to be involved in the efficient organization of chlorophyll molecules during light interception and therefore may be of importance to the efficient operation of the photosynthetic apparatus.

Lodgepole pine is an economically important North American forest crop species and has been well studied at the isozyme and gene level. My research was concentrated on characterizing genomic copies of the gene because I was interested in the structure of the surrounding gene region. There is strong evidence that there are approximately eight *cab* gene copies for the Type I gene product and several more for the Type II gene product in the LHCII of angiosperms. The observation that there are numerous copies, for both Type I and Type II *cab* genes within a species means that any study interested in the evolution and variation of the different copies has to be concerned with ensuring that all copies examined are from a single individual rather than variants of a single copy carried in different individuals. Therefore it seemed logical to focus my attention on the genetic information available from a single individual through the construction and screening a genomic library from a single tree.

The objectives of this thesis were as follows:

1. To construct a genomic library from a single individual of lodgepole pine and screen with a homologous probe for *cab* genes.

2. To sequence and characterize the identified *cab* gene(s) from lodgepole pine and to contrast it to other published information concerning *cab* gene structure, regulation and evolution.

3. To compare and contrast the deduced transit peptide deduced sequences from all known PSII *cab* genes with the hope of identifying any common features.

4. To analyze phylogenetically the nucleotide sequence of the coding region of all known *cab* PSII genes.

Chapter 2 covers construction of a genomic library from a single individual of lodgepole pine, the screening of the library and the identification of positive λclones. Chapter 3 covers the sequencing and characterization of a single Type I *cab* gene. The fourth chapter compares the proposed amino acid sequence of the transit peptides from the published *cab* genes with the objective of identifying any similarities between and within the different transit peptide types. In chapter 5 I have used cladistic and phenetic methods to establish the phylogenetic relationships between the coding regions of the Type I and Type II genes in PSII.

LITERATURE CITED

Bobola, M.S., R.T. Eckert and A.S. Klein. 1992. Restriction fragment variation in the nuclear ribosomal DNA repeat within and between *Picea rubens* and *Picea mariana*. Can J. For. Res. 22:255-263.

Demmin, D.S., E.J. Stockinger, Y.C. Chang and L.L. Walling. 1989. Phylogenetic relationships between the chlorophyll a/b binding protein (CAB) multigene family: an intra- and interspecies study. J. Mol. Evol. 29:266-279.

Green, B.R., E. Pichersky and K Kloppstech. 1991. Chlorophyll a/b-binding proteins: an extended family. TIBS 16:181-186.

Jansson, S., and P. Gustafsson. 1990. Type I and Type II genes for the chlorophyll a/b-binding protein in the gymnosperm *Pinus sylvestris* (Scots pine): cDNA cloning and sequence analysis. Plant Mol. Biol. 14:287-296.

Karlin-Neumann, G.A., B.D. Kohorn, J.P. Thornber and E.M. Tobin. 1985. A chlorophyll a/b-protein encoded by a gene containing an intron with characteristics of a transposable element. J. Mol. Appl. Genet. 3:45-61.

Keegstra, K. 1989. Transport and routing of proteins into chloroplast. Cell 56:247-253.

Kohorn, B.D., E. Harel, P.R. Chitnis, J.P. Thornber and E.M. Tobin. 1986. Functional and mutational analysis of the light-harvesting chlorophyll a/b protein of thylakoid membranes. J. Cell Biol. 102:972-981.

Kuhlbrandt, W. and D.N. Wang. 1991. Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. Nature 350:130-134.

Neale,D.B., N.C. Wheeler and R.W. Allard. 1986. Paternal inheritance of chloroplast DNA in Douglas-Fir. Can. J. For. Res. 16:1152-1154.

Pain, D., Y.S. Kanwar and G. Blobel. 1988. Identification of a receptor for protein import into chloroplasts and its localization to envelope contact zones. Nature 331:232-237.

Rake, A.V., J.P. Miksche, R.B. Hall and K.M. Hansen. 1980. DNA reassociation kinetics of four conifers. Can. J. Genet. Cytol. 22:69-79.

Strauss, S.H., J.D. Palmer, G.T. Howe and A.H. Doerksen. 1988. Chloroplast gemones of two conifers lack a large inverted repeat and are extensively rearranged. Proc. Natl. Acad. Sci. 85:3898-3902.

CHAPTER TWO

SEARCHING FOR CAB GENES IN LODGEPOLE PINE

INTRODUCTION

Plants and bacteria that undergo photosynthesis have evolved a group of light harvesting proteins which are coded for by the chlorophyll a/b (*cab*) genes. These proteins form complexes which act to intercept incoming light energy and to transfer this photon energy to the organism's photosystems. Numerous angiosperm species have been studied in the past decade and the information from these studies has resulted in the identification of several general features for the *cab* genes. These include the organization of the *cab* genes in a small nuclear multigene family in which the gene products are transported and routed into the chloroplasts. The genes are divided into groups based on their location and function in the photosystems. Regulation of the *cab* genes has been shown to be tissue specific and light dependent and that several regulatory elements control this expression.

The study of *cab* genes has intensified in the last several years, and gymnosperm species are proving to have unique features. For example, it has been shown that some conifers have the ability to produce chlorophyll in the dark (Bogdanovic 1973, Schou 1951). This is in contrast to angiosperms where little or no chlorophyll is synthesized in the dark and any observed chlorophyll is quickly degraded. In flowering plants the expression of the *cab* genes has been shown to be light and tissue dependent (Simpson et al. 1986). Nevertheless, it has recently been demonstrated that substantial amounts of

the CAB protein was present in both dark and light grown cotyledon tissue in *Pinus thunbergii* (Yamamoto et al. 1991) and Douglas-fir (Alosi et al. 1990). It is interesting to note that this observation is not conifer specific, because in larch (deciduous conifer) *cab* genes are expressed in light-grown but not in dark-grown seedlings (Hutchison et al. 1990a). Larch, therefore, exhibits responses similar to the situation observed in angiosperms.

Genetic studies of conifers have shown that they have extremely large nuclear genomes ranging in size from 1.86 X 10^{10} (red pine) to 5.7 X 10^{9} (white cedar) bp per haploid nucleus (Miksche and Dhillon, 1981). In contrast the nuclear genome of white ash, a woody angiosperm, has been estimated at one sixth the size of white cedar (Miksche and Dhillon, 1981). The size of the nuclear genome of lodgepole pine has been estimated at 1.1 X 10^{10} bp. The reason for such a large genome size in conifers is unknown; however, it is known that the majority of this DNA is repetitive. Studies of the reassociation kinetics for jack pine, a close relative of lodgepole pine, indicate that only 20% of its large genome contains unique sequences (Rake et al. 1980). This has two consequences; first it is difficult to construct genomic libraries that are sure to contain 99% of the DNA, and second, with such a large component of repetitive DNA it is difficult to locate single-copy genes. Nevertheless, there has been an increase in the number of genes located and studied in conifers. For example, the *actin* gene in lodgepole pine (Kenny et al. 1988), alcohol dehydrogenase (*adh*) in radiata pine (Kinlaw et al., 1988) some of the *cab* genes from scots pine (Jansson and Gustafsson 1990), *Pinus thunbergii* (Yamamoto 1988a) and larch (Hutchison et al. unpublished) and the small

subunit for ribulose bisphosphate carboxylase (*rbs*) in larch (Hutchison et al. 1990b) and *Pinus thunbergii* (Yamamoto et al. 1988b) have recently been described.

MATERIALS AND METHODS

DNA Extraction and Restriction Digestion

Total cellular DNA from conifers for use in Southern analysis was isolated according to the procedure described by Murray and Thompson (1980) and modified by Rich Jorgenson (Advanced Genetic Systems, Oakland, CA). Ten grams of fresh needle tissue was ground in liquid N_2 until the largest pieces were less than 0.5 cm. The powdered tissue was homogenized in extraction buffer (20 ml TE, 40 ml 1.75M sorbitol, 2 ml 10% BSA, 0.2 ml 2-mercaptoethanol and 20 g PEG 3350; bring up to 200 ml with water) for 10 seconds with a polytron. The homogenate was filtered through cheesecloth and miracloth and the nuclei and organelles were pelleted at 9000 rpm for 15 minutes at 4^0C. The dark green pellet was resuspended, with difficulty, using a Number 9 paintbrush in CT wash buffer (0.5 ml solution B [0.5 M tris, 0.25 M EDTA], 1.0 ml 1.75 M sorbitol, and 5 µl 2-mercaptoethanol; bring up to 5 ml with H_20). The resuspended pellet was treated with 1 ml of N-lauroylsarocine (5%) mixed and incubated for 45 minutes at room temperature. One ml of 5M NaCl was added and the suspension was mixed by inverting gently. 0.8 ml of 10% CTAB (cetyldimethylethyl ammonium bromide) was added and the solution was inverted several times and then incubated at 60^0C for 10 minutes. Chloroform was added and the extract was mixed slowly until an emulsion formed. The aqueous layer was transferred to a 30 ml corex tube and 10 ml of ice cold isopropanol

was added. The solution was mixed and stored at -20°C for 10 minutes. The precipitated DNA was then hooked out of the solution and transferred to a tube containing 5 ml 76% ethanol/10mM ammonium acetate and left for more than 20 minutes at 4°C. The DNA was resuspended in 0.5 ml of TE. Lodgepole pine genomic DNA used for library construction was processed according to Watson and Thompson (1986).

Total genomic DNA of each species was digested with 8 U of either *Eco*RI, *Hin*dIII or *Bam*HI for five hours at 37°C and the fragments were separated on a 0.5% agarose gel run at 60 volts for 16 hours.

Hybridization conditions

Gels were blotted onto Zetabind (CUMO) nylon membrane and probed sequentially with the following nick-translated probes: the complete lambda lodgepole pine clone, PNcAB10-1, containing a Type I *cab* gene; the *cab* cDNA, pCab3, from petunia (Dunsmuir et al., 1983); and the genomic *cab* clone, pLgAB19, from duckweed (Karlin-Neumann et al., 1985). Probes were nick-translated, and hybridizations occurred at 65°C for 37 or 65 hours for the conifer genomic blots and overnight (16 hours) for the cloned DNAs. Genomic hybridization times were calculated based on probe size, probe concentration and volume of hybridization solution (Maniatis et al. 1982). The larger the probe or the larger the volume of hybridization solution the longer the hybridization time. Genomic blots were washed under conditions of low stringency (once at room temperature in 2XSSC followed by two washes of 2XSSC + 1% SDS at 42°C for 30 minutes each). Autoradiography was performed at -70°C with intensifying screens for times varying from

20 hours to two weeks. All other filters were washed twice at room temperature for 15 minutes each in 2XSSC, followed by twice at 65°C for 30 minutes in 2XSSC + 1% SDS, followed by two washes at 65°C in 2XSSC + 0.1%SDS for 30 minutes each. Autoradiography was carried out at -70°C with intensifying screens overnight.

Library Construction

Lodgepole pine DNA used for library construction was partially digested with 0.6 U of *Sau*3A for one hour at 37°C to produce a range of fragment sizes with *Bam*HI compatible ends. The fragments were separated on a 10-40% continuous sucrose gradient, and fractions containing fragments between 15-23 kb were collected for library construction.

Preparation of Lambda Arms

The replacement vector λdash (Stratagene) was digested with *Bam*HI to release the two arms from the central fragment. The terminal *Bam*HI site on each arm is compatible for the *Sau*3A ends produced from partial digestion of the genomic DNA. The central fragment was further subjected to *Xho*I digestion to produce ends that were not compatible with the left and right arms, thereby decreasing the chance of recovering nonrecombinant phage. The central fragment was physically removed by adjusting the vector DNA preparation to 10 mM $MgCl_2$ and incubating the sample for one hour at 42°C (Kaiser and Murray, 1985), thereby promoting annealing of the arms at the *cos* site. The

arms were separated from the central fragment by sucrose gradient fractionation, and the collected fractions were checked on a 0.6% agarose gel.

Preparation of Packaging Extracts

Extracts for the packaging of phage DNA were prepared by one of the following methods. Initially, complementing extracts were prepared in which each extract supplied required proteins not found in the other extract. For example, the sonicated extract was produced from the lysogenic strain BHB2690. The strain allows for the formation of λ preheads but is defective for gene *D* which is involved in the packaging of λ DNA. The second lysogenic strain was BHB2688 which supplied the packaging protein *D*. Extracts were prepared following the protocols outlined in Arber et al. 1982. Each extract was stored separately in 15 µl and 10 µl aliquots at -70^{0}C.

The second method utilized the single nonsuppressing lysogenic strain, SMR10, to replace the two strain packaging extracts. SMR10 was prepared as described in Rosenberg 1987.

Tests on the packaging efficiencies of these extracts suggested that the most efficient combination was the addition of the BHB2688 freeze-thaw extract to the SMR10 extracts.

In vitro *Packaging*

Packaging extracts were thawed on ice for ten minutes. The freeze-thaw extract was added to the sonicated extract and mixed gently. The DNA was added and the

mixture was incubated at room temperature for one hour. This mixture was then diluted by the addition of one ml of Phage Suspension Buffer (PSB; 10mM tris, 100mM NaCl, 10mM $MgCl_2$ and 0.05% gelatin). The library number was tested by adding one µl of the library to 300 µl of plating culture. The λ was allowed to be absorbed for 20 minutes at 37°C. Three mls of top agar (50°C) was added to the plating mixture, which was then poured onto LB plates and incubated at 37°C overnight.

Bacterial Cultures

Plating cultures for λ work were prepared from one of the following bacterial strains: LE392, P2392, NM621 or DL538. These strains were grown in 10 ml LB with the addition of 0.2% maltose overnight. The maltose induces the maltose operon in the bacteria that encodes the *lam*B gene of the λ receptor, with the result that *E. coli* cells absorb λ more efficiently. The next day a 10 ml 1/50 dilution was grown at 37°C for 6-7 hours. The culture was spun down at 5000 rpm for ten minutes at room temperature. The cell pellet was resuspended in 40% of the original volume of LB with the addition of 10 mM $MgSO_4$.

The success of library construction was checked by plating ligation aliquots on bacterial strains utilizing the *spi*⁻ phenotype. The most common of these pairs of strains was LE392 and P2392. Both recombinants and non-recombinants grew on LE392, while only clones in which the central 15 kb of λ had been replaced by foreign DNA grew when plated on P2392, because the P2392 strain is a P2 lysogen and therefore does not permit any phage expressing the *gam/red* genes to survive. These genes are located

within the central region of λ which is replaced by foreign DNA during library construction. This means that it is possible to estimate the efficiency of the library by comparing the titre on each strain when using the same concentration of DNA. Efficiency of recombinant libraries ranged up to 75%.

Construction of plant genomic libraries is difficult in part because of the inefficient recovery of recombinant clones on traditional bacterial strains. It has been established that there are at least four loci in *E. coli* K12 that encode functions capable of degrading cloned eukaryotic DNA. These include the *hsd*, *mcrA*, *mcrB* and *mrr* loci. Of these functions, the *mcr*A and *mcr*B nucleases recognize specific sequences only when they are methylated. Plant DNAs have high levels of 5-methylcytosine and the effects are therefore very important in cloning plant DNA. The use of *mcr*- hosts can overcome this problem. In this study I used NM621 (*hsd*R, *mcr*A, *mcr*B, *rec*D), which had been shown to improve the yield of recombinant human DNA tenfold to give cloning efficiencies of $>10^7$ pfu/μg of human DNA (Whittaker et al., 1988). Use of the strain NM621 increased the yield of recombinant clones by 50 times.

Competent cells were prepared using DH5α and DH5αF' cell lines following the protocols of Hanahan (1985).

Preparation of Cloned DNA

The replicative form of M13 was isolated according to the alkaline lysis method (Sambrook et al. 1990). M13 clones were transfected into freshly prepared DH5αF' cells

where the F' episome allowed the formation of pili on the outer cell wall thereby permitting infection of the filamentous phage.

Well isolated white plaques were grown in 5 mls of a 1:50 dilution of a fresh overnight culture of DH5αF' cells with vigourous shaking for six hours. The extruded virus particles were collected by centrifugation following treatment with 20% PEG and 2.5M NaCl. The viral particles were disrupted by vortexing the resuspended pellet in phenol. The M13 DNA was precipitated and resuspended in 20 µl of TE and 3-5 µl were used in a sequencing reaction. All sequencing reactions were initiated with the universal primer.

Double-stranded plasmid and phagemid DNA were isolated from mini-preparations using the alkaline lysis procedure (Sambrook et al. 1990). Large scale preparations of phage or plasmid DNA were purified on CsCl gradients. Isolation of single-stranded DNA from the phagemids pTZ18/19R was achieved by super infection with the addition of 5×10^9 helper phage, R408, to a multiplicity of infection (MOI) of 10. The template quality was improved greatly by infecting the phagemid culture to a MOI of 10 with 6.4×10^9 of the helper phage M13K07 at a titre of 5×10^{10} pfu/ml. After one hour the non-infected cells were selected against by the addition of 10 µl of a kanamycin solution (70mg/ml). Single-stranded DNA was isolated following the Pharmacia T7 sequencing kit protocol. Phage were precipitated by the addition of 0.25 volumes of 3.5 M ammonium acetate/20% PEG and incubated on ice for 30 minutes. Following precipitation the pellet was resuspended in 400 µl of TE. An equal volume of phenol/$CHCl_3$ was added and the mixture was vortexed for 30 seconds. The upper

aqueous layer was removed following centrifugation for one minute. This phenol/$CHCl_3$ extraction was repeated once and followed by a chloroform extraction. The DNA was precipitated by addition of 0.1 volume of 3M sodium acetate and two volumes of ethanol. The sample was incubated on dry ice for 10 minutes and pelleted by a 10 minute centrifugation. The pellet was washed once with 70% ethanol, air dried for 10 minutes and then resuspended in 20 µl of H_2O.

Construction of Deletions

Supercoiled phagemid DNA was converted to open circles by digestion with DNase I in the presence of a saturating amount of ethidium bromide. The nick was widened to a small gap by a short treatment with exonuclease III. The single-stranded nuclease activity of *Bal*31 was utilized to cleave the single strand bridging the gap. The result was a single random gap in each molecule. The sample was then digested with a unique restriction endonuclease near the primer. This action resulted in the release of the fragment between the restriction endonuclease site and the *Bal*31 generated gap. The resulting mixture should contain a complete size range of molecules. Klenow was added to fill in 5' overhangs and when the sample was run on a gel a smear was observed. The desired sizes were sliced out of the gel and the DNA extracted with Geneclean (Bio101, San Diego, CA). The different sized deletions were blunt end ligated overnight at 22^0C. The re-ligated samples were transformed into DH5α competent cells and plated on LB^{amp} plates with incubation at 37^0C overnight.

Deletions in M13 clones were prepared following unidirectional digestion with exonuclease III (Henikoff 1987). Unfortunately the success of this protocol was marginal and many of the clones did not exhibit the anticipated sequential deletions. Much better success for the pTZ18/19R subclones was achieved by replacing the S1 nuclease with 15U of mung bean nuclease for each time point. Approximately 300 bases were cleaved off the clone per minute of incubation at 37^{0}C. A 1.5 kb clone was easily and completely sequenced using six deleted sub-clones. There was variation within each time point such that the last 600 bases were sequenced using three different sub-clones from the same time point.

Sequencing Single-stranded DNA

Sequencing reactions were carried out following the protocols outlined in the Sequenase or Pharmacia sequencing manuals. The only changes were the use of 2.5 µCi instead of 10 µCi of α-^{35}SdATP per reaction and the use of the reverse primer only.

Sequencing Double-stranded DNA

Excellent double-stranded sequencing was achieved by denaturing 2 µg of plasmid DNA, obtained from a plasmid alkaline lysis mini-preparation, with the addition of 2 µl of 2M NaOH and incubating at 80^{0}C for 5 minutes. The denatured template was precipitated by the addition of 11 µl 7.5M ammonium acetate and 90 µl of 99% ethanol. The mixture was incubated at -70^{0}C for 10 minutes and the DNA was collected by centrifugation for 20 minutes at 4^{0}C. The pellet was washed with 80% ethanol and then

air dried for 10 minutes at room temperature. The pellet was dissolved in 10 µl of H_2O. The annealling and sequencing reactions were as outlined in the sequencing manuals in which both the universal and reverse primer were used. The only change was the use of 10 µCi of α-^{32}PdATP per reaction.

Sequencing Gels

All sequencing gels were 6% polyacrylamide, run at 65 watts for 3 hours for short reactions or at 700 volts for 16 hours for long reactions. Most gels were run with 1 X TBE in the top chamber and 1M sodium acetate and 1 X TBE in the lower chamber. This produces a simple salt gradient which increases the amount of readable sequence at the bottom of the gel.

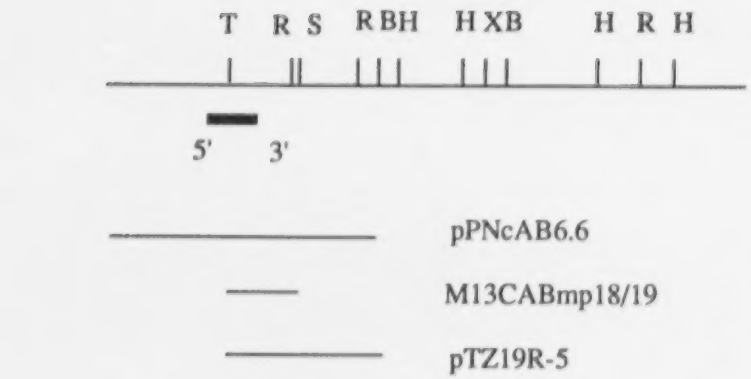
RESULTS

An amplified lodgepole pine genomic library with an estimated titre of $1.2x10^6$ pfu/ml (J. Kenny, Personal Communication) was grown on twelve plates with approximately 50,000 phage per plate. A 1.9 kb genomic clone of a Type II *cab* gene, pLgAB19, from duckweed (*Lemna gibba*) was used as the heterologous probe to screen for clones that shared homology. Seven primary positives were selected from this screening and verified by secondary screening with freshly nick-translated pLgAB19. Fourteen positives from the secondary screening were selected for further analysis. Ten of these clones were digested with *Sal*I to release the inserts from the EMBL3 arms and the fragments were separated on a 0.5% agarose gel. Of the ten putative positives, there

were four unique *Sal*I restriction patterns, three other clones exhibited digest patterns already observed and no DNA was retrieved from the final three clones. The four unique clones were tested by performing a Southern using the partial *cab* cDNA, pCab3, to ensure hybridization. The pCab3 probe is 500 bp long and contains 411 bp of coding region for a Type I *cab* gene and 100 bp of 3' flanking region. A single clone (λPNcAB10-1) produced a strong hybridization signal. Because I was interested in obtaining a single homologous probe for the purpose of screening my own library, further analysis was carried out on this single clone and the other clones were abandoned.

λPNcAB10-1 and pTZ19R-5

Restriction map information of the clone λPNcAB10-1 (Fig.2.1), and Southern hybridizations, suggested that a single 6.6 kb *Bam*HI fragment contained the complete *cab* coding region. This fragment which was adjacent to the left arm of the vector was cloned into pUC19 (pPNcAB6.6) and characterized by Southern hybridization. The region of strongest hybridization to the cDNA Type I *cab* gene from petunia, pCab3, was contained in a 1.8 kb *Sst*I-*Sal*I fragment. This fragment was cloned in both directions into M13mp19 and nested deletions were prepared. Initial sequence data failed to correspond to any published *cab* sequences. This lack of homology and problems with erratic deletions prompted me to go back and perform some fine scale restriction mapping on the region of interest. Therefore using an internal *Sst*I site which bisected the *cab* gene in pPNcAB6.6 and the *Sst*I site in the pUC polylinker a 4 kb fragment was released from the *Bam* clone. This fragment was cloned into the phagemid pTZ19R (new clone



Legend

cab gene

1.0 kb

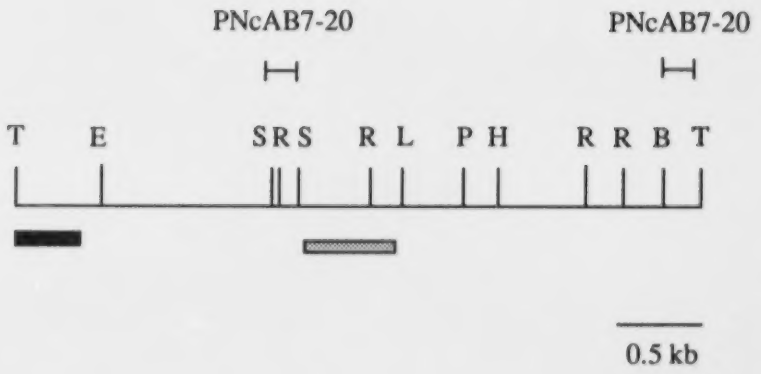
B *Bam*HI
H *Hin*dIII
R *Eco*RI
S *Sal*I
T *Sst*I
X *Xba*I

Figure 2.1. Restriction map of the lambda clone PNcAB10-1. The lines under the map indicate the fragments sub-cloned. The *Sst*I (T) site was added after it was identified in the sub-clones. The 3' end of the gene was identified by sequence analysis.

pTZ19R-5; Fig.2.2). Because both the *cab* cDNA (pCab3), and genomic (pLgAB19) clones hybridized to this fragment, it was used as a homologous probe for my genomic library from lodgepole pine. Further sequence analysis of pTZ19R-5 confirmed that this was indeed a partial copy of a Type I *cab* gene from PS II when compared to published genes. Although the complete copy of the gene was within the clone pPNcAB6.6, the *Sst*I site used for subcloning is found at position 432 of the coding region in effect dividing the gene in half. The partial cDNA probe from petunia (pCab3), used to confirm the presence of a *cab* gene, covers the coding region from bases 415 to 835 thereby identifying only the 3' end of the gene.

Preliminary Southern analysis (data not shown) suggested that there were at most only a couple of closely related *cab* genes within the genome of lodgepole pine. This is in contrast to angiosperms who have been shown to contain up to 16 copies of the Type I *cab* genes within their genomes. These seemingly contradictory results were intriguing. To try to resolve this question it would be necessary to construct my own genomic library from a single individual of lodgepole pine to ensure that any identified copies were different members of the same gene family rather than multiple copies of the same allele.

pTZ19R-5 was nick-translated with 150 μCi of $\alpha^{32}P$ and used to screen my genomic library. Ten positives were selected from the primary screening of this library. Secondary screening suggested that several of these recombinant phage contained DNA homologous to the probe. Each of the clones was digested with *Sal*I and the fragments separated on a 0.6% gel. This gel was blotted and probed, first with the 4 Kb homologous fragment and later with the *cab* cDNA. *Sal*I digestions suggested that there



LEGEND

Region of hybridization to cab clone

Region of hybridization of the 400bp *Bam*HI-*Sst*I fragment from clone 4-10
see pg.29

B BamHI
H HindIII
L SalI
P SphI
R EcoRI
E EcoRV
S SmaI
T SstI

Fig. 2.2. Restriction map of pTZ19R-5. Lines above the restriction map indicate the regions of hybridization for the 1.7 kb *Bam*HI/*Sst*I fragment from clone PNcAB7-20. Several EcoRI sites were added, after fine mapping, to this smaller subclone of PNcAB10-1.

were eight different clones picked from the library. The other two clones represented duplicates. Southern analysis using the 4 kb homologous probe suggested that five of the unique clones were hybridizing (Fig.2.3). Hybridization with the *cab* cDNA from petunia indicated hybridization to one clone, λPNcAB2-10, after 10 hours (Fig.2.4). λPNcAB2-10 was further characterized and found to contain a Type I *cab* gene (Fig.2.5; see next chapter and Fig.3.1 for description). Longer exposure times suggested that the cDNA may hybridize weakly to all the clones. These results suggested that there was enough evidence to pursue and characterize these clones.

DESCRIPTION OF THE λPNcAB CLONES

λPNcAB4-10

The restriction map for clone 4-10 (Fig.2.6) suggested that the region of hybridization to pTZ19R-5 is found across two *Bam*HI fragments. Unfortunately, when this clone was probed with the complete *cab* gene from lodgepole pine clone 2-10 (PNcAB2; see Fig.2.5) there was no indication of hybridization. Therefore whatever hybridized to this clone from within pTZ19R-5 does not appear to have been a *cab* sequence.

Southern analysis of λ4-10 probed with pTZ19R-5 indicated that there was a homologous 400 bp *Bam*HI-*Sst*I fragment within 4-10 (see Fig.2.13). This 400 bp fragment was isolated and used as a probe against the other putative lodgepole pine *cab* clones. Unfortunately, it did not hybridize to any other clone. This result suggests that, although this sequence is homologous to pTZ19R-5, it is not similar to any sequence

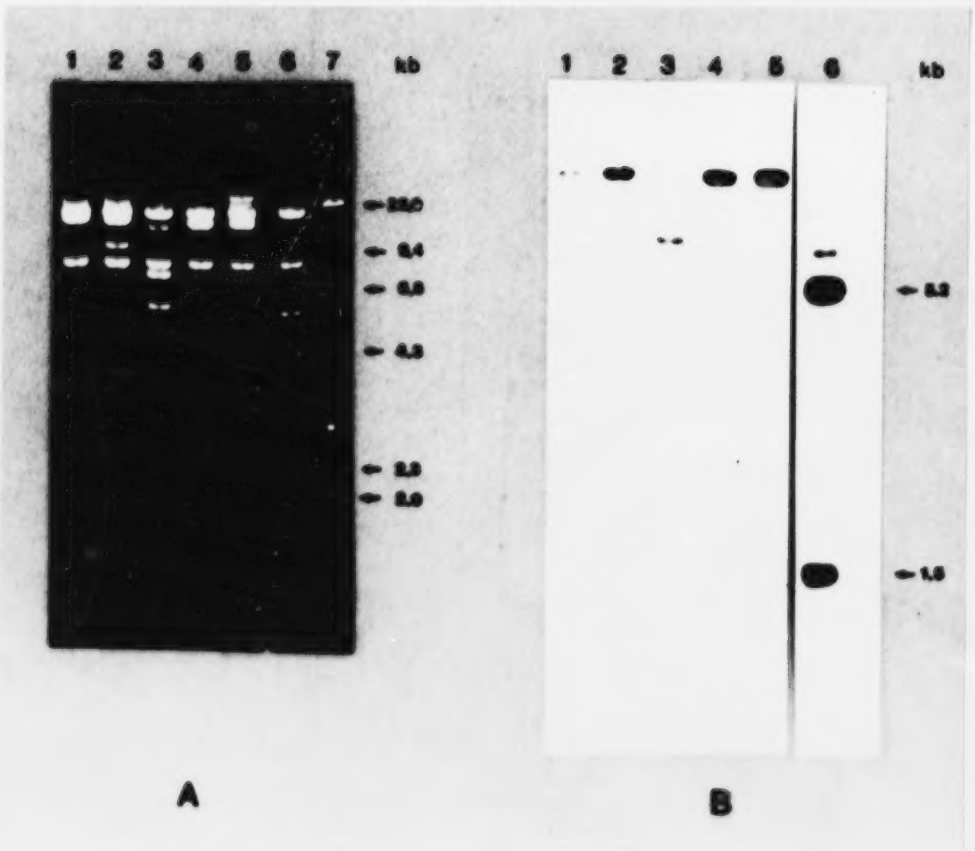
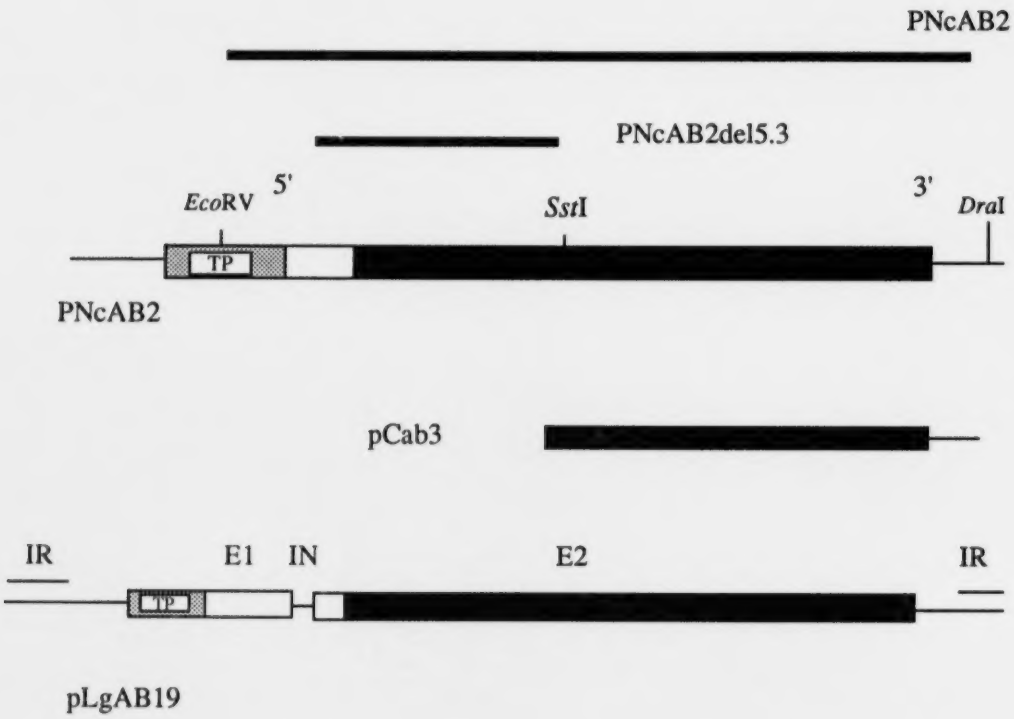


Figure2.3. *A*. Agarose gel stained with ethidium bromide illustrating the digestion of lodgepole pine (PNcAB) lambda clones with *Sal*I. Only the clones discussed within the text are shown. Lane 1 represents clone 7-20, lane 2 is 7-10, lane 3 is 5-10, lane 4 is 4-10, lane 5 is 4A, lane 6 is 2-10 and lane 7 is lambda DNA digested with *Hin*dIII and *Pst*I. *B*. The Southern of this gel probed with the lodgepole pine homologous probe pTZ19R-5. Fragment sizes are listed on the side.

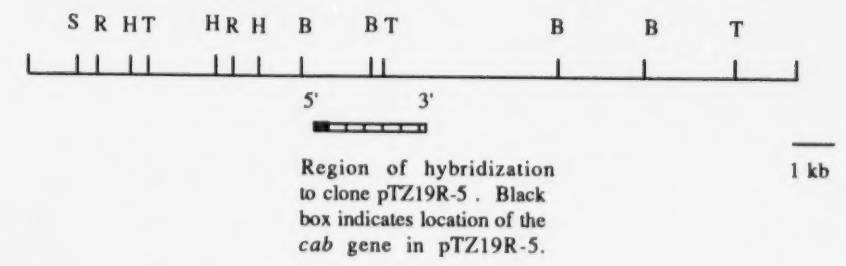
Figure 2.4. Southern of the agarose gel shown in Fig. 2.5. probed with the *cab* cDNA probe pCab3 from petunia. Lanes 1 through 6 are lodgepole pine lambda clones 7-20, 7-10, 5-10, 4-10, 4A and 2-10 respectively. The autoradiogram was developed after 10 hours at -70C.



Legend

TP Transit peptide
IR Inverted repeats
E1 Exon 1
E2 Exon 2
IN Intron

Figure 2.5. Map of the lodgepole pine Type I *cab* gene, PNcAB2 from the lambda clone PNcAB2-10. Lines above the map indicate the homologous *cab* probes described in the text. The boxes below the map indicate the two heterologous probes used in this study. The shaded areas indicate regions of sequence similarity. The hatched box indicates the transit peptide for clone PNcAB2.



Region of hybridization to clone pTZ19R-5 . Black box indicates location of the *cab* gene in pTZ19R-5.

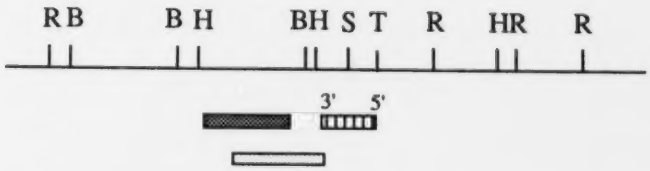
B *Bam*HI
H *Hin*dIII
R *Eco*RI
S *Sal*I
T *Sst*I

Figure 2.6. Restriction map of lambda clone 4-10.

within any of the other λclones. Therefore it may just be the result of a random region of similarity found in pTZ19R-5 and not shared with the other clones. Or it may be that this particular fragment is just outside the regions cloned into the λvector. When this 400 bp fragment is hybridized back against pTZ19R-5 it hybridizes to a region approximately 2 kb downstream of the *cab* gene in pTZ19R-5 (see Fig.2.2).

λPNcAB5-10

Clone 5-10 hybridized strongly to the homologous probe, pTZ19R-5 and to the 1 kb *cab* (PNcAB2) coding region from lodgepole pine. Restriction digest analysis (Fig.2.7) suggested that the region of hybridization was in the centre of the clone. Although there did not appear to be any hybridization between this clone and the *cab* cDNA pCab3 from petunia, when a blot of single and double digests was probed with the Type II *cab* clone pLgAB19 from duckweed there was a strong signal. It is interesting to note that there are different regions of hybridization between the lodgepole pine (pTZ19R-5) and duckweed (pLgAB19) probes (see Fig. 2.7). Because the Type I *cab* gene, PNcAB2, and the Type II *cab* clone, pLgAB19, hybridize to different fragments within the λclone 5-10 this suggests that this clone may contain a single copy of a Type I and a Type II *cab* gene. Unfortunately there is no evidence that the two probes used, pLgAB19 and the coding region from PNcAB2, are that specific. All previous Southern information suggests that both probes would cross hybridize and either one would hybridize to both Type I and Type II *cab* genes.



1 kb

LEGEND

Region of hybridization of PNcAB2

Region of hybridization of pLgAB19

Region of hybridization of pTZ19R-5

B BamHI

H HindIII

R EcoRI

S SalI

T SstI

Figure 2.7. Restriction map of lambda clone 5-10

λPNcAB7-10

Clone 7-10 hybridized to pTZ19R-5 strongly and to the genomic *cab* gene pLgAB19 from duckweed. As observed with λ5-10 the two clones hybridize to two adjacent regions of the clone (Fig.2.8). Subcloning and partial sequencing of the 1.7 Kb *Bam*HI fragment adjacent to the left arm confirmed that clone 7-10 contained a truncated copy of the 3' end of a Type I *cab* gene (Fig.2.9). Over this short sequence there were only six differences between this clone and the lodgepole pine *cab* gene PNcAB2. This suggests one of several possibilities. 1) that this is just a copy of the same gene, 2) that it is the complementary allele for the same gene, or 3) that it is a very closely related member of the same gene family. As with λ5-10 I am unable to offer a credible explanation for the differential hybridization patterns of pLgAB19 and pTZ19R-5. However, the sequence information supports the Southern data produced by probe pTZ19R-5 and PNcAB2.

λPNcAB7-20

pTZ19R-5 hybridized to a 1.7 kb *Bam*HI-*Sst*I fragment located within the centre of the clone (Fig.2.10). In contrast to clones 5-10 and 7-10, clone 7-20 did not hybridize to either the duckweed *cab* (pLgAB19) or the full *cab* gene from clone 2-10 (PNcAB2). The 1.7 kb fragment was subcloned into pTZ19R (pTZ7-20:1.5kb) and the ends were sequenced with the universal and reverse primers. Sequence comparison indicated that the sequence of 7-20 was 75% similar to a short stretch of 200 bp at the end opposite to

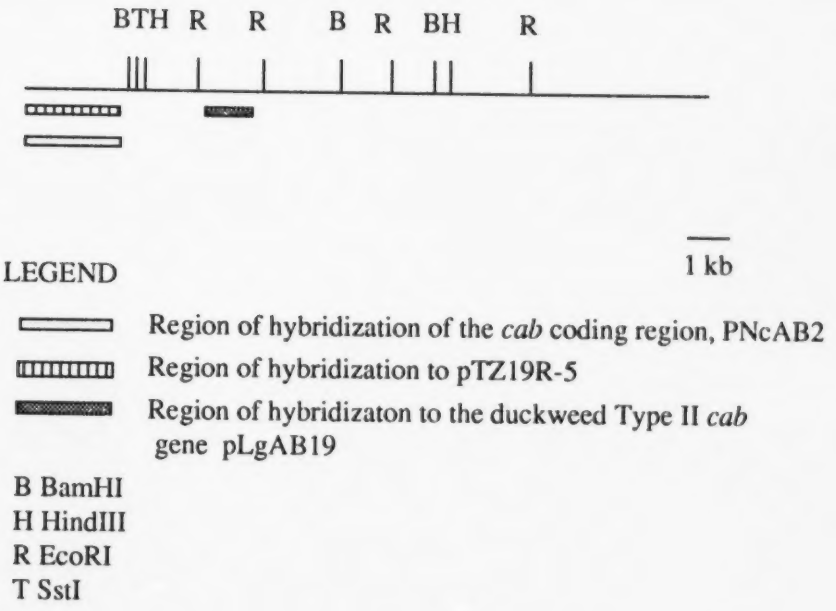
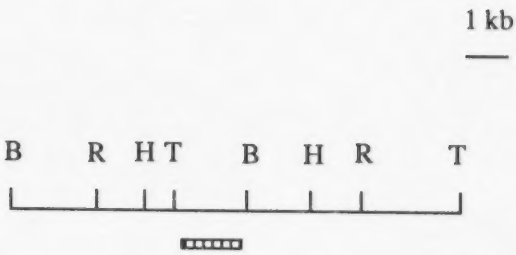

BTH R R B R BH R
1 kb
LEGEND
Region of hybridization of the cab coding region, PNcAB2
Region of hybridization to pTZ19R-5
Region of hybridizaton to the duckweed Type II cab gene pLgAB19
B BamHI
H HindIII
R EcoRI
T SstI

Figure 2.8. The restriction map of lambda clone 7-10.

```
λPNcAB7-10   TTTTCAGCGAAGGGAGGGCTTGACTACCTGGGAAACCCTAGC
PNcAB2       TTTTCAGCGA-GGGAGGGCTTGACTACCTGGGAAACCCTAGC

λPNcAB7-10   TTGGTTCACGCGC**GAGA**CATTCTTGCAATCTGGGCATGC-AGG
PNcAB2       TTGGTTCACGCGCAGAGCATTCTTGCAATCTGGGCATGCCAGG

λPNcAB7-10   TCATCCTCATGGGCGCCGTGGAGGGCTACCGTATTGCGGGCGG
PNcAB2       TCATCCTCATGGGCGCCGTGGAGGGCTACCGTATTGCGGGCGG

λPNcAB7-10   TCCTCTGGGGAGAGGTGACCGACCCCATCTACCCCGGG--CAGCT
PNcAB2       TCCTCTGGG-AGAGGTGACCGACCCCATCTACCCCGGGGGCAGCT

λPNcAB7-10   TCGACCCGCTGG-ACTTGCGGAC**G**ACCCCGACGCTTTCGCAGAG
PNcAB2       TCGACCCGCTGGGACTTGCGGACTACCCCGACGCTTTCGCAGAG

λPNcAB7-10   CTGAAGGTGAAGGAGATCA-GA-CGGGCGGT-GG-CATGTTTT
PNcAB2       CTGAAGGTGAAGGAGATCAAGAACGGGCGGTTGGCCATGTTTT

λPNcAB7-10   **T**CATGTTCGGAT
PNcAB2       CCATGTTCGGATTCTTCGTGCAGGCCATCGTCACCGGGAAA
```

Figure 2.9. Comparison of the sequence of λPNcAB7-10 and the cab gene PNcAB2 from lodgepole pine. Bold letters indicate nucleotide changes between the sequences. Spaces have been inserted to improve alignment.



LEGEND

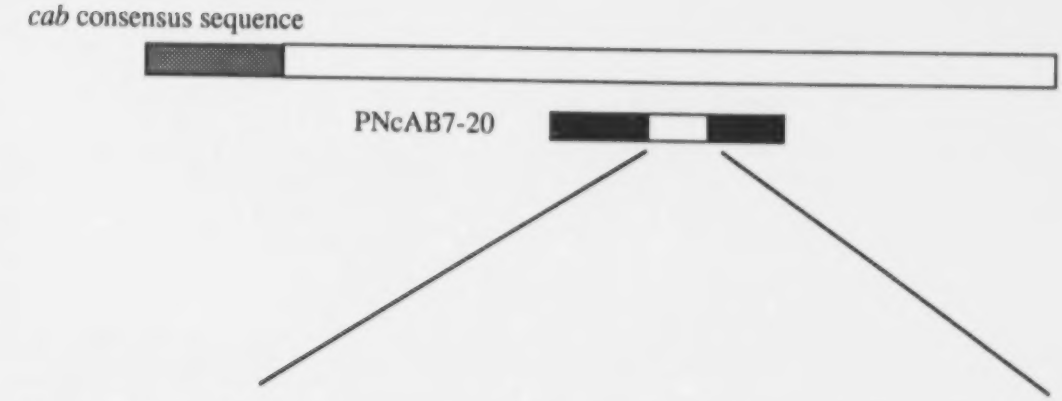
Region of hybridization for pTZ19R-5

B *Bam*HI
H *Hin*dIII
R *Eco*RI
S *Sal*I
T *Sst*I

Figure 2.10. Restriction map of the lambda clone PNcAB7-20.

the *cab* gene in clone pTZ19R-5. It is interesting to note that an internal 40 bp sequence used as a probe to search the EMBL sequence library is very similar to several cab genes (Fig.2.11). It is also worth noting that the matches are all either pine or monocotyledonous species.

When this 1.7 kb fragment (pTZ7-20:1.5kb) was used to probe back against digests of pTZ19R-5 an interesting phenomenon was observed. This clone hybridized to two different areas on pTZ19R-5 (see Fig. 2.2). This strange observation was supported by the sequence data. For example, as stated above when the 1.7 kb fragment was sequenced with the reverse primer the sequence was 75% similar to the sequence at the 3' end of the subclone pTZ19R-5. In contrast, sequence data using the universal primer showed 56% homology with a short 90 bp region at the SmaI/EcoRI site approximately 1.5 kb from the 5' end of pTZ19R-5. The distance between these two regions, approximately 2 kb, is close enough that errors in the estimation of fragment sizes could explain these observations. This would mean that this is a 1.7 kb homologous region shared by pTZ19R-5 and pTZ7-20:1.5. However, the Southern data does not support this explanation of shared similarity across the total 1.7 kb fragment. I conclude therefore that pTZ19R-5 contains fragments of unknown size which share some similarity to unconnected regions in the clone λclone 7-20. This could mean that pTZ19R-5 contains a repeated element and that this element is found in some of my clones or it is a transposable element as suggested for pLgAB19 (Karlin-Neumann et al. 1985). When the two sequences from clone pTZ7-20:1.5 are compared against each other there is 35%




```
pINEab13        CCCATCTACCCAGGGGGCAGCTTCGACCCGCTGGGACTTCGCGGAA
LGLHCPI         ..GC.G.....C..C.....................G..GGC..ACG
BARII-III       ..TCA.CC.GGC..CCAGTA...............C...-..C..C
RICE CAB2       ..GC.......G..C...GC.............C..C..-...C..T
LGAB19          ..GC.......C..C..GGC..............G..GGC..ACG
λPNcAB7-20      A.TTAAACTGAG.CTACA...ATACTAT....C..........C..
pINEab9         ..T.....T..C..C..T.A.........A........G-..A..A
CRErbcS         TCCGGTTCCTGTCCGTGTCTTTGTCCCT..T.CACC........C.

pINEab13        GACCCTGAGGCTTTCGCGGAGCTGAAGGTGAAGGAGATCAAGAACG
LGLHCPI         -....C........-.................................
BAR II-III      .......TTA.C...........T.....................T.
RICE CAB2       .....G.....C...........C.......................
LGAB19          -....G..T.................A.....................
λPNcAB7-20      -C...........ACACCTA..GGTCT.G.C..CAC.CAAGAA.G
pINEab9         .....G..TG................A.....................
CRErbcS         AC..T.C.TC.CC.T..TTGCAG.CT..ATCCCCT.CCTGGAGTT.
```

Figure 2.11. Comparison of the sequence from clone 7-20 which shares a 42 nucleotide region with other *cab* genes and a single *rbcS* gene from *Chlamydomonas reinhardtii*. The 42 nucleotides of interest are in bold letters. The *cab* genes are from scots pine (pINEab13 and pINEab9), duckweed both Type I (LGLHCPI) and Type II (LGAB19), barley Type III gene from PSII and rice *cab* Type II. Clone λPNcAB7-20 is in bold lettering. Open box indicates conserved *cab* coding sequence, black box non-homologous sequence and hatched box represents the transit peptide sequence.

similarity between them. This does not rule out the repetitive fragment theory because these regions may represent sequence outside the repeat region.

Based on sequencing, λPNcAB2-10 and λPNcAB7-10 contain *cab* coding regions. PNcAB2-10 was a complete copy of a Type I gene and PNcAB7-10 a partial copy of a Type I gene. Clone PNcAB5-10 hybridized to both the 1 kb probe (PNcAB2) of the *cab* coding region from lodgepole pine and the genomic clone from duckweed (pLgAB19). However short 200 bp regions of sequence from λ5-10 did not match published *cab* gene sequences. Therefore a conserved 282 bp fragment from position 150 to 432 of the coding region of this lodgepole pine *cab* gene (PNcAB2) was used as a hybridization probe (PNcAB2:del5-3)in a Southern analysis of the other lodgepole pine lambda clones. Surprisingly, the probe only hybridized to the lambda clone (λPNcAB2-10) from which it was isolated (Fig. 2.12). This was puzzling because of the other hybridization evidence (not shown) that suggested that λ5-10 contained DNA homologous to the DNA from the genomic *cab* clone of duckweed. Since the sequence of this probe (PNcAB:del5-3) is very similar to all other published Type I and II *cab* genes from PSII, it was expected to hybridize to any lodgepole pine copies of these genes. When these same digested samples were probed with pTZ19R-5, there was evidence of hybridization in all five clones within the hour (Fig.2.13).

The unusual results reported here may be explained by one of three possibilities. First, it is possible that the clone, pTZ19R-5 contains a region of repetitive DNA that is dispersed throughout the lodgepole pine genome. This would explain why the clones λ4-10 and λ7-20 hybridize to pTZ19R-5 but not to the *cab* probes, PNcAB2, pCab3 or

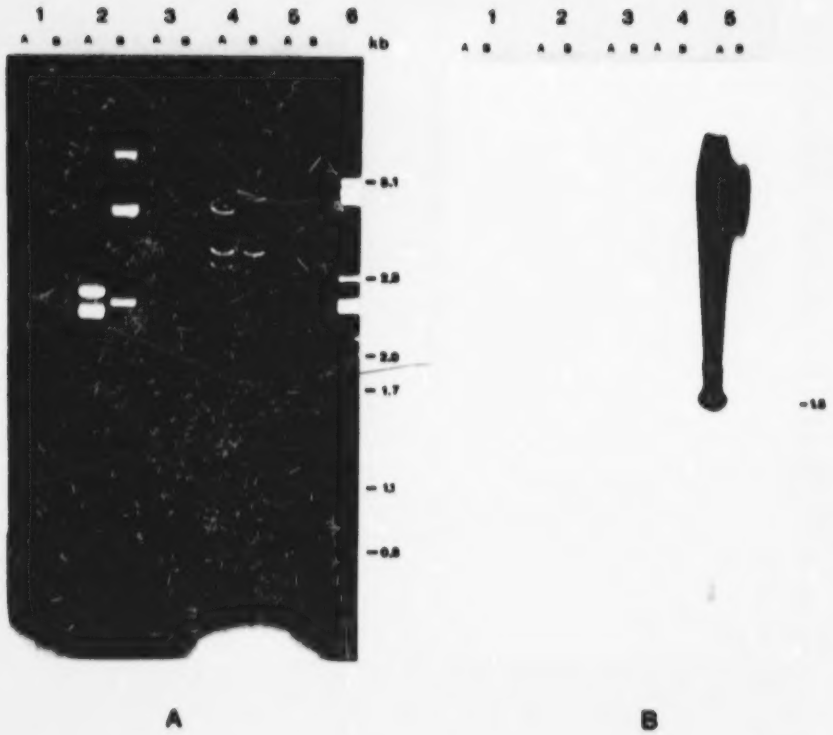


Figure 2.12. *A*. Agarose gel of the lodgepole pine clones digested with either *Bam*HI (lanes labelled with B) or *Sst*I/*Bam*HI (lanes labelled with A). The samples labelled 1 through 6 are clones 7-20, 7-10, 5-10, 4-10, 2-10 and lambda DNA digested with *Pst*I. *B*. The Southern hybridization of the same gel, probed with a 282 bp fragment from clone 2-10.

Figure 2.13. *A*. Autoradiogram of the same gel as shown in Fig.2.12. probed with pTZ19R-5 and autoradiographed at -70 for one hour. *B*. Longer exposure of the same gel. Autoradiogram left for 12 hours at -70C. Lane 1 is clone 7-20, lane 2 is 7-10, 3 is 5-10, 4 is 4-10, 5 is 2-10 and 6 is lambda DNA digested with *Pst*I. Selected fragment sizes are labelled.

pLgAB19. A similar phenomenon has been reported in tomato (Bernatzky et al. 1988). A 1.65 kb fragment (CR1) located within the 3 kb spacer between two *cab* genes was found to be 75% A/T rich and to be flanked by imperfect repeats. Southern hybridization of digested genomic DNA from tomato indicated that this fragment hybridized to 20-30 fragments. Thirteen of these fragments were mapped to seven of the tomato's 12 chromosomes. As well, it is interesting to note that only two of the repeated fragments mapped to CAB loci. The other fragments appeared to be randomly dispersed throughout the genome. Unfortunately, comparison of this sequence to the various sequences from the lodgepole pine clones did not reveal any similarity.

The second possibility is that pTZ19R-5 contains a transposon-like sequence that hybridizes to similar elements found within the lodgepole pine genome. An interesting feature of pLgAB19, the clone used to select pTZ19R-5, is the occurrence of large (150 bp) inverted repeats which flank the transcription unit (Karlin-Neumann et al. 1985). As well, the intron sequence found within this gene exhibits features of a transposable element including short (7bp) direct repeats at the intron junction so that when the intron is spliced out only one copy of the repeat is left. These direct repeats immediately flank imperfect inverted repeats in which 10 of the 15 bases at the 5' end of the intron are complementary to those at the 3' end.

It is possible that the pLgAB19 hybridized to some sequence found in pTZ19R-5 and that this sequence, if it occurs, is hybridizing to similar transposable sites throughout the genome.

The third possibility is that these other λclones contain more distantly related cab genes possibly from PSI or related to CP24 or CP29.

DISCUSSION

Construction and screening of my genomic library was based on the estimation that approximately 3.5 X 10^6 recombinants would be required to have a 99% chance of locating a single copy gene. My best library had approximately 4 X 10^5 recombinants. In theory there is a less than a 25% chance of locating a single copy gene with such a small representation of the lodgepole pine genome. Fortunately I was able to locate one complete copy, one partial copy and perhaps one distantly related copy of a *cab* gene. Therefore the isolation of three putative *cab* genes was unexpected; however, it may be due to the large repetitive DNA component within conifers which allowed for this better than expected success. It has been noted that cloned repetitive eukaryotic DNA is often unstable in bacterial strains (Whittaker et al. 1988). It may be that I inadvertantly selected against clones carrying repetitive DNA by transfecting bacteria that caused the restriction and degradation of those λ clones containing repeated DNA. If this was the case it suggests that the lodgepole pine genome is composed of regions of repetitive DNA interspersed with islands that contain less repetitive DNA or coding regions. This bacterial restriction system might allow for the production of λ clones containing approximately 17 kb inserts containing single copy genes.

One of the most interesting aspects of studying *cab* genes in conifers was to determine the number of copies of the *cab* genes in lodgepole pine. It has been

demonstrated that the number varies within the angiosperms and the most recent estimates put the number of Type I genes for PSII at 8-16 copies and 1-2 copies of the Type II genes. This question has been addressed indirectly for gymnosperms. No Southerns of digested genomic DNA have been produced to confirm this number in gymnosperms. The incomplete picture which is arising is that conifers may have many fewer copies of Type I genes than do the angiosperms. For example, two complete copies of Type I cab genes and a single partial Type II *cab* gene from a scots pine cDNA library and a single complete Type II *cab* gene from a *Pinus thunbergii* cDNA library have been reported. As well, initial reports suggest that there are only three *cab* genes in *P. palustris* and two in the gymnosperm *Ginkgo biloba* (Chinn et al. 1992) This suggests that the number of copies in gymnosperms/conifers are low.

In this present study I have screened several conifer species with probes derived from lodgepole pine to assess *cab* copy number (data not shown). Several points are worth noting. First, I found little or no evidence of cross hybridization among several species of conifer when using a lodgepole pine probe (i.e. λPNcAB10-1 or pTZ19R-5). For example, there was hybridization for the lodgepole pine and closely related jack pine samples; however, no evidence of hybridization for white or ponderosa pine (*Pinus strobus, P. ponderosa*), or any of the spruces, firs, larch or Douglas fir (*Picea mariana, P. glauca, P. pungens, Abies balsamea, A. grandis, A. lasiocarpa, Larix laricina and Pseudotsuga menziesii*). This is interesting because in angiosperms the *cab* genes are well conserved and a *cab* gene from pea has cross hybridized to such species as tomato,

petunia, duckweed and was also used to screen the scots pine library. Perhaps my hybridization conditions were too stringent for a general survey of the conifers.

Secondly, the hybridization study in lodgepole pine also revealed that there may not be as many copies as might be expected based on angiosperm studies. When a 1.8 kb probe containing the last 403 bp of the coding region was used to probe lodgepole and jack pine samples there appeared to be four bands of hybridization (data not shown). The 403 bases included in the probe are well conserved among species and between Type I and Type II cab genes. Therefore, it was expected to hybridize to all LHCII *cab* gene copies. This Southern analysis evidence along with the location of only three putative copies of *cab* genes in lodgepole pine and finally the publication of only two Type I cDNA copies from scots pine suggests that perhaps in conifers or pines that the number of Type I *cab* genes is less than is found on average in angiosperms.

The clone λPNcAB10-1 selected from a genomic lodgepole pine library contained a copy of a Type I *cab* gene as determined by partial sequencing of the subclone pTZ19R-5. This homologous lodgepole pine clone has been shown to contain the last 403 nucleotides of the *cab* coding region. This segment, along with 3.5 kb of downstream flanking region was used as a probe to screen my lodgepole pine library. Of the five λ clones selected for analysis, only one was unequivocally a complete copy of a *cab* gene. This clone, λPNcAB2-10, is characterized in the next chapter. Another clone (λ7-10) contained a truncated fragment of a Type I *cab* gene and a third clone (λ5-10) hybridized to the lodgepole pine *cab* genes (pTZ19R-5 and PNcAB2) and a Type II *cab* gene probe from duckweed. What is worrisome is that clone 5-10 did not exhibit any

signs of hybridization when probed with a 282 bp fragment from the 5' end of the coding region from clone 2-10. This suggests that this clone may be a truncated copy missing the 5' end or that the probe sequence may not be within clone 5-10, although since the region of hybridization mapped to the centre of the λ clone, a complete copy was expected. As well I was unable to locate any sequence which was similar to other published cab gene copies except for some strong similarity to a very short region of 20 nucleotides from the 3' flanking region from two Type II cab genes.

The other two clones (λ4-10 and λ7-20) contained regions which hybridized to pTZ19R-5 but not to the *cab* gene probes (PNcAB2, PNcABdel5.3, pLgAB19 and pCab3).

It could be that each of the undetermined clones contain some sequence that is found in pTZ19R-5 but not contained in any of the other clones. That is they each contain some random, unrelated similarity to pTZ19R-5. This may be supported by the observation that when pTZ19R-5 hybridized to a 400 bp fragment in 4-10 and when this fragment was used as a probe against the other clones it did not hybridize to any other clone except for the one it came from, 4-10 and to pTZ19R-5.

I am left with several unresolved questions including what is it in the clone pTZ19R-5 that hybridizes to the λ clones? Are the λ clones related or just hybridizing to random pieces of clone pTZ19R-5? It seems likely that the best strategy would have been to create nested deletions through the subclone pTZ19R-5, to sequence these deletions and then to search Genbank or EMBL with the complete sequence from pTZ19R-5 in hopes of locating some similar sequences. This strategy would also have

permitted for very selective hybridization of the λ clones by using the 300 bp deletions as probes for identifying regions of hybridization.

LITERATURE CITED

Alosi, M.C., D.B. Neale and C.S. Kinlaw. 1990. Expression of *cab* genes in Douglas-fir is not strongly regulated by light. Plant Physiol. 93:829-832.

Arber, W., L. Enquist, B. Hohn, N.E. Murray and K. Murray. 1983. Experimental methods for use with lambda. In *Lambda II* (ed. R.W. Hendrix,J.W. Roberts, F.W. Stahl and R.A. Weisberg), pp.433-466. Cold Spring Harbour Laboratory, Cold Spring Harbour, New York.

Bogdanovic, M. 1973. Chlorophyll formation in the dark. I. Chlorophyll in pine seedlings. Physiol. Plant. 29:17-18.

Chinn, E.E., J. Silverthorne and W. Peer. 1992. Regulation of chlorophyll a/b binding protein genes in *Ginkgo biloba*. Abstract in Supplemental Issue Plant Physiol. :99.

Dunsmuir, P., S.M. Smith and J. Bedbrook. 1983. The major chlorophyll a/b binding protein of petunia is composed of several polypeptides encoded by a number of distinct nuclear genes. J. Mol. Appl. Genet. 2:285-300.

Hanahan, D. 1985. Techniques for transformation of E. coli. In *DNA Cloning Volume I: A Practical Approach*. (ed. D.M. Glover), pp.109-136. IRL Press Ltd, Oxford

Henikoff, S. 1987. Unidirectional digestion with exonuclease III in DNA sequence analysis. Methods Enzymol. 155:156-165.

Hutchinson, K.W., C.D. Sherman, J. Weber, S.S. Smith, P.B. Singer and M.S. Greenwood. 1990a. Maturation in larch. II. Effects of age on photosynthesis and gene expression in developing foliage. Plant Physiol. 94:1308-1315.

Hutchison, K.W., P.D. Harvie, P.B. Singer A.F. Brummer and M.S. Greenwood. 1990b. Nucleotide sequence of the small subunit of ribulose-1,5-bisphosphate carboxylase from the conifer *Larix laricina*. Plant Mol. Biol. 14:281-284.

Jansson, S. and P. Gustafsson. 1990. Type I and Type II genes for the chlorophyll a/b-binding protein in the gymnosperm *Pinus sylvestris* (Scots pine): cDNA cloning and sequence analysis. Plant Mol. Biol. 14:287-296.

Kaiser, K. and N.E. Murray. 1985. The use of phage lambda replacement vectors in the construction of representative genomic DNA libraries. In *DNA Cloning Volume I: A Practical Approach.* (ed. D.M. Glover), pp.109-136. IRL Press Ltd, Oxford

Karlin-Neumann, G.A., B.D. Kohorn, J.P. Thornber and E.M. Tobin. 1985. A chlorophyll a/b-protein encoded by a gene containing an intron with characteristics of a transposable element. J. Mol. Appl. Genet. 3:45-61.

Kenny, J.R., B.P. Dancik, L.Z. Florence and F.E. Nargang. 1988. Nucleotide sequence of the carboxy-terminal portion of the lodgepole pine actin gene. Can. J. For. Res. 18:1595-1602.

Kinlaw, C.S., D.E. Harry, D.D. Sleeter and R.R. Sederoff. 1988. Use of heterologous probes to isolate and characterize conifer genes. in *Molecular Genetics of Forest Trees.* W.M. Cheliak and A.C. Yapa (eds). Petawawa National Forestry Institute Information Report PI-X-80, 95pp.

Miksche, J.P. and S.S. Dhillon. 1981. Nucleic acid research techniques in forest genetics. In: *Advances in Forest Genetics* (India) pp.329-371.

Murray, M.G. and W.F. Thompson. 1980. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8:4321

Rake, A.V., J.P. Miksche, R.B. Hall and K.M. Hansen. 1980. DNA reassociation kinetics of four conifers. Can. J. Genet. Cytol. 22:69-79.

Rosenberg, S.M. 1987. Improved *in vitro* packaging of λ DNA. Methods Enzymol. 153:95

Simpson, J., M. Van Montagu and L. Herrera-Estrella. 1986. Photosynthesis-associated gene families: differences in response to tissue-specific and environmental factors. Science 233:34-38.

Sambrook, J., E.F. Fritsch and T. Maniatis. 1989. *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbour Press, Cold Spring Harbour, N.Y.

Schou, L. 1951. On chlorophyll formation in the dark in excised embryos of *Pinus jeffreyi.* Physiol. Plant. 4:617-620.

Watson, J.C., and W.F. Thompson. 1986. Purification and restriction endonuclease analysis of plant nuclear DNA. Methods Enzymol. 118:57-74.

Whittaker, P.A., A.J.B. Campbell, E.M. Southern and N.E. Murray. 1988. Enhanced recovery and restriction mapping of DNA fragments cloned in a new λ vector. Nucleic Acids Res. 16:6725-6736.

Yamamoto, N., Y. Mukai, M. Matsuoka, Y. Kano-Murakami, Y. Tanaka, Y. Ohaski, Y. Oseki and K. Odani. 1991. Light-independent expression of *cab* and *rbcS* genes in dark-grown seedlings. Plant Physiol. 95:379-383.

Yamamoto, N., M. Matsuoka, Y. Kano-Murakami, Y. Tanaka and Y. Ohashi. 1988a. Nucleotide sequence of a full length cDNA clone of light harvesting chlorophyll a/b binding protein gene from green dark-grown pine (*Pinus thunbergii*) seedling. Nucleic Acids Res. 16:11829.

Yamamoto, N., Y. Kano-murakami, M. Matsuoka, Y. Ohashi and Y. Tanaka. 1988b. Nucleotide sequence of a full length cDNA clone of ribulose bisphosphate carboxylase small subunit from green dark-grown pine (*Pinus thunbergii*) seedling. Nucleic Acids Res. 16:11830.

CHAPTER THREE

MOLECULAR CHARACTERIZATION OF A SINGLE COPY OF THE CAB GENE IN LODGEPOLE PINE

INTRODUCTION

In green tissue of plants, the efficient absorption and transfer of light energy into the photosystems (PSI and PSII) requires that chlorophyll molecules be bound together tightly and in a defined orientation. Each photosystem has its own light harvesting protein complex (i.e. LHC-II) to control this requirement. The binding of chlorophyll molecules within the LHC is achieved by the production of the chlorophyll a/b binding protein. These chloroplast proteins are encoded by nuclear genes (*cab* genes) and include information for a 33 - 44 amino acid transit peptide that allows the mature peptide to be transported into the chloroplast. This transit peptide is then cleaved off releasing a mature peptide of 230 - 235 amino acids of 25,000-27,000 MW. Within PSII the LHC-II has been identified as a trimer composed of three identical Type I CAB monomers (Kuhlbrandt 1984). Each monomer consists of a single polypeptide and fifteen associated chlorophyll molecules (eight chlorophyll *a*, seven chlorophyll *b* and several carotenoids; Butler and Kuhlbrandt, 1988).

Studies of the *cab* genes have revealed a complex multigene family organized in a hierarchical system in which each photosystem has at least two types of *cab* genes (reviewed by Buetow et al., 1988). In petunia it has been determined that the 16 copies of the Type I *cab* genes in photosystem II may be divided into 5 subfamilies based on

the homology of their 5' and 3' untranslated regions (Dunsmuir 1985). As well there is at least one Type II gene from PSII (Stayton et al. 1986).

In tomato, eight genes have been identified as Type I genes of PSII (Pichersky et al., 1985) and two more are Type II genes in PSII (Pichersky et al., 1987a). As well, tomato has been shown to have Types I, II, and III genes in PSI (Pichersky et al. 1987b; Pichersky et al., 1988; Pichersky et al., 1989).

Most LHCs are associated with PSII (Green et al., 1991), therefore it would seem that PSII coding genes are of more importance. Of these, the Type I genes are more numerous, are without introns and are often clustered in the genome (Dunsmuir et al. 1983, Dunsmuir 1985, Leutwiler et al. 1986, Pichersky et al. 1985). Although the Type I genes diverge in nucleotide sequence from Type II genes, the gene structure is strongly conserved between both types (Pichersky et al. 1987a,b).

There is a paucity of information on nuclear encoded genes in conifer species. Of these, the only genomic DNA information available is from a larch *rbc*S gene (Hutchinson et al., 1990) and the carboxy portion of the lodgepole pine actin gene (Kenny et al. 1988). cDNAs are known for *cab* genes (Jansson and Gustaffson 1990, Yamamoto et al., 1988b) in *Pinus sylvestris* and *P. thunbergii*; and *rbc*S in larch (Hutchinson et al. 1990) and in pine (Yamamoto et al., 1988a). In all cases these clones have been isolated from libraries constructed from pooled individuals. If the *cab* genes within pines are organized in a fashion similiar to that of angiosperms then it is important to reduce the possibility of selecting multiple copies of the same gene family member by using genomic DNA extracted from single individuals. This would ensure that any *cab* gene identified

was one of the two possible complementary copies of each member found in a single individual and not the same copy from different individuals. Here I report the isolation and sequence of a single copy of a *cab* gene from a genomic library constructed from a single lodgepole pine.

MATERIALS AND METHODS

Construction of Packaging Extract and Lambda Arms

Packaging extracts were prepared as described in Sambrook et al. (1989). The sonication extract was prepared using SMR10 (Rosenberg 1985, 1987), a nonsuppressing lysogen of *E. coli* C. The freeze/thaw lysate was constructed from inducible BHB2688 (Hohn 1979).

The bacteriophage vector λ Dash (Stratagene) was used for the production of arms. Twenty-five μg of λ Dash was digested with 50 U *Xho*1 at 37°C for two hours to release the central fragment. Complete digestion was verified on a minigel. The vector was double digested by adjusting the reaction mixture to 100 mM NaCl by adding 50 mM NaCl and 50 U of *Bam*H1 to create a complementary cloning site for *Sau*3A ends. The restricted vector was extracted twice with phenol/$CHCl_3$ and precipitated with 0.5 volumes of 7M ammonium acetate and 2.5 volumes of 95% ethanol.

Successful double digestion was checked by religating the vector and plating out on LE392.

Extraction of Lodgepole Pine Genomic DNA

Ten grams of fresh needle tissue from a single specimen of *Pinus contorta* var. *latifolia* were ground in liquid N_2 and processed according to Watson and Thompson (1986). DNA was purified through two CsCl centrifugations and the size was compared to intact lambda DNA (48 kb) on an agarose gel ensuring that the conifer DNA was larger than 50 kb. 50 µg of total genomic DNA was partially digested with 0.6 U of *Sau*3A for one hour at 37°C to produce fragments between 15 and 23 kb containing ends compatible with the *Bam*HI site of the vector arms. This reaction was loaded onto a 10% - 40% continuous sucrose gradient and centrifuged for 17 hours at 20°C in a SW40 rotor at 25,000 rpm. Fractions were collected in 500 µl aliquots. Every third fraction was precipitated with 0.5 vol of 7.5M ammonium acetate and 1 vol. of isopropanol and resuspended in 10 ul of TE. These fractions were analyzed on a 0.5% horizontal agarose gel in 1% TBE. Fractions containing fragments in the 15-23kb range were pooled and precipitated. The pellet was resuspended in 10 µl of TE.

Ligation Reaction

Approximately 0.4 µg of genomic DNA from the sucrose gradients described previously was ligated to 1.6 µg of vector arms so there was a total of 2 µg in a ratio of 4:1 vector to insert. This was incubated at 15°C for 7 hours and then an aliquot of ligated sample was checked against an unligated control on a 0.5% TBE minigel. One µl of a 10^{-2} dilution of the packaging mixture was plated on LE392 and P2392 bacterial strains to evaluate the number of recombinants formed. Test conditions suggested that 77% of

the phages were true recombinants. When another aliquot of the same dilution was plated on NM621 there was an observed increase of 50X the number of recombinants.

Library Construction

A 100 µl aliquot of the library was incubated with 300 µl of NM621 plating culture and incubated for 25 minutes at 37°C. This was added to 7 ml of liquid LB top agar (LTA+Mg), mixed and poured onto a large 150mm petri dish. This was repeated for eight plates. These plates were incubated at 37°C until plaques were visible, the plates were removed before the plaques had grown together to prevent recombination. Eight plates were prepared to produce an estimated library of $4x10^5$ recombinants of which approximately 77% were true recombinants. Double lifts of each of the eight plates were taken using large circular nitrocellulose filters as the transfer medium. The 16 filters were denatured, neutralized and dried according to Sambrook et al. (1989).

Hybridization and Identification of Positives

All the filters were hybridized together in a box with approximately 150 mls of prehybridization solution (Westneat et al., 1988) at 42°C with agitation. A 4.0 kb homologous probe containing the *cab* gene was nick-translated (Rigby et al.,1977), added to the prehybridization solution and incubated at 42°C for 40 hours with agitation. The filters were washed once in 2XSSC for 2 minutes at room temperature, once in 2XSSC for 30 minutes at 42°C and finally twice in 2XSSC, 0.1% SDS for 45 minutes at 42°C. All washes were performed with some agitation. The filters were placed on film (Kodak)

with intensifying screens for 4 days at -70°C. Suspected positives were picked with the large end of a pasteur pipette. The agar plug was blown into an eppendorf tube containing 500 µl of Phage Suspension Buffer and the phage were allowed to diffuse at room temperature for 2.5 hours. An aliquot was diluted 10^{-2} of which 2 µl was incubated with 200 µl of NM621 plating culture for 20 minutes at 37°C and then plated on small (90mm) petrie dishes. These plates allowed secondary screening of potential positives and increased the ease of selecting individuals instead of clusters.

Growth of Phage and DNA Extraction

Identified positives were picked with the small end of the pasteur pipette and the plug was blown into 10ml of LB with 10mM $MgCl_2$, and 200 µl of plating culture and incubated at 37°C at 250 rpm. Lysis usually occurred within 5 hours of initation of incubation. Chloroform was added to prevent any further protein synthesis and the bacterial debris was pelleted. The lysate was transferred to another tube and the phage particles were precipitated by the addition of PEG and NaCl. The phage particles were ruptured by extracting with phenol, then phenol/$CHCl_3$ and finally chloroform. The DNA was precipitated by the addition of 0.5 volumes of 7.5M ammonium acetate and 2.5 volumes of 95% ethanol.

Restriction Mapping

Each restriction digest contained 5 µl of DNA and was digested with 1 µl (10 U) of the appropriate restriction endonuclease in a total volume of 15 µl at 37°C for 1.5

hours. The samples that were double digested were modified by increasing the NaCl concentration and the addition of the correct restriction endonuclease and incubated for a further two hours. The RNA was removed by the addition of 1 µg of DNase-free RNase A (10 mg/ml) to each digestion and allowing them to sit at room temperature for 20 minutes. Fragments from the single and double digests were separated on 0.7% agarose gels in 1%TBE. The DNA was visualized by staining the gels in 300 mls of ethidium bromide (0.5 µg/ml) solution for 20 minutes with agitation. The gels were treated according to the protocol for GeneScreen Plus and the DNA was transferred overnight. The filters were hybridized to the homologous probe pTZ19R-5 according to the conditions outlined above.

Sequencing

Fragments containing the suspected *cab* gene were cloned into the complementary phagemid vectors pTZ18/19R containing the reverse primer site (Mead et al., 1986). Deletions were created in one of two ways, either by the method outlined in Barnes et al., (1983) or by a modified Henikoff (1987) method substituting mung bean nuclease for S1 nuclease. Growth and isolation of the single stranded phage is outlined in the Pharmacia Sequencing Kit.

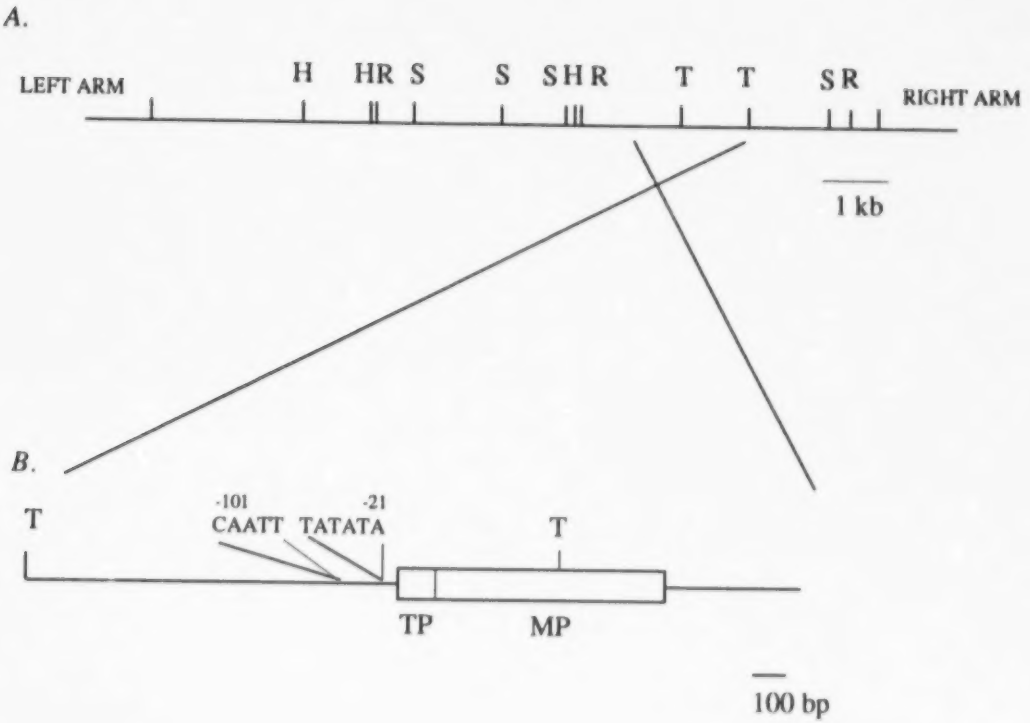
Sequence data were compiled and managed with the IBI MacVector programs and the Genbank BioNet Online services.

RESULTS

cab *Coding Region*

A genomic library constructed of lodgepole pine seedlings (gift of J. Kenny) was screened with pLgAB19 (gift of E. Tobin), a genomic clone of a LHCII Type II *cab* gene from duckweed (*Lemna gibba*). A single recombinant phage containing a *cab* copy was isolated. This copy was cloned into the phagemid pTZ19R and was used to screen a genomic library constructed from a single individual of lodgepole pine. Eight recombinant phage were selected following hybridization to the probe. Restriction digests and Southern analysis using pCAB3, a partial cDNA clone of a Type I *cab* gene from petunia (gift of P. Dunsmuir) suggested that only one of the clones contained a copy of a *cab* gene.

Clone PNcAB2 has been sequenced over a 2.5 kb region. This sequence includes 1.1 kb of upstream flanking region, 60 nt of the untranslated 5' end, 825 nt of coding sequence, 115 nt of 3' untranslated region and 400 nt of the downstream flanking region (Fig. 3.1). The coding region is comprised of a single open reading frame of 822 nucleotides encoding an apoprotein with a suspected 40 amino acid residue transit peptide and a mature CAB peptide of 234 amino acids (Fig. 3.2). There is no evidence of the occurrence of introns within the coding region. Based on sequence similarity between other *cab* genes this clone appears to be a copy of a Type I *cab* from PSII. Sequence comparison of *cab* genes in the EMBL library found that the two scots pine and a Japanese pear tree (Kano-Murakami, unpublished 1991) contained *cab* genes most similar to PNcAB2, followed by a *Lemna* Type I *cab* gene (Kohorn et al. 1986), maize



TP Transit peptide
MP Mature peptide
H *Hin*dIII
R *Eco*RI
S *Sal*I
T *Sst*I

Figure 3.1. The organization of the *cab* gene, PNcAB2 in lodgepole pine. *A*. The restriction map of the λ clone. *B*. The structure of the *cab* gene and surrounding area. The numbers give the position of the CAAT and TATA boxes in relation to the initiation codon.

```
ACACACATAGTTAATATCTAGAACTAAACAAATATAATGAACAACTAACTCTTTAGATAA  60
TTATATATCTTCTATTTATTAACAATATACATATAATTCGACAATCAATAATTAAAAAAA
AGAAGGTAAACTTTTGAATTAAAAAAAAAAAGGCATAGTGTAACTTATAGATGAAATGGA  180
CTTTTCACAGGTAGCCAACAATACAATTGCAAGATTGCAATTGTTTGTTGTTTAAAATAC
ATTTTTGAATGCATAAATGGATTGCATTTTATAGAGTAAACAACTGAAAATCAAGTTTTT  300
AAAAATTTCTCCATGATTTTGATTAAAAATAACGTGAAAAATGCTCTAAAAAAAGACCAT
ATCATCTATAAGTTAAACTAAGTCTTAAAAATATATGTATATATAAAATTAAATGCTAAA  420
AAATACCATTTCATCGATAAATTAACTCTTTTCCAAATTAAAATGGTTGTTGCGCAGAGA
GAGAAAGTTGATAGGGGCATTCTAAAGCAGAGTGTCAAAGTTGAAAATTTGTTGCCAAAA  540
AAAGAGAGTTAAATAACGATACAAAACATTTAAATAATATAGGTTCATAAAAAAAATAAT
GACCATATAAATAAATTCTAGGATGGGTAAATTTTTTTAGAACTTAGATTTTGAAAAAAA  660
ATTACACCTAAAATAATTTGATCTCTGACCCATAATAAATTCAAAATTAAATGGATCTAA
CCTGATCATTAAACCAAGAGAGATTGGATCTAACCGAACATGATAAATCCAAAATTAAAT  780
AGATATATACTAATCATATAACCCAACATTAACGCTAAACAAGGTTACCGGCAATAAGAA
GGCTTAACCATTACATATAGTGATGGATGATATTTTTTCTTTTAATGGTGCGTGGGCAAT  900
CCGGTTCCACCCAATAGAGGATGAGGCTGAGATTGCATGACACAAGGATATGGTTTGAAT
GTACGAAGATGATGTGGGAGGGCAATTGTGGGTGTGATTGGCCGAGGAGATTCGGAAACG 1020
ATAGGAGATGAAATTATTTGGTCTGCGGTTGCTCTATATATTGAGCTGATGCAGTTCAAG
GCCAGAAGAGAACTGGTACAAGCTAAACTAATCTTCTGTTTGTGCGCCTACTGCTGCAAC 1140
AATGGCAAGCTGCGGGATCGGGTCGTGCGGTTTCGCTGGGGGACAGATATCATCGCTGAA
  M  A  S  C  G  I  G  S  C  G  F  A  G  G  Q  I  S  S  L  K>
GCCTCACACCAACCAGTTGCTGGGAGTCGGCGCGGGGGTTCATGGCGAGGCGCGAGTGAC 1260
  P  H  T  N  Q  L  L  G  V  G  A  G  V  H  G  E  A  R  V  T>
GATGAGGAAGGCGACCACCAAAAAAGTGTCAGTTTCAGCGAGCACATCGCCATGGTATGG
  M  R  K  A  T  T  K  K  V  S  V  S  A  S  T  S  P  W  Y  G>
GCCGGATCGGGTTCTCTACTTGGGGCCTTTCTCCGGCGAGCCGCCATCTTACTTGACCGG 1380
  P  D  R  V  L  Y  L  G  P  F  S  G  E  P  P  S  Y  L  T  G>
TGAGTTCCCCGGTGACTACGGGTGGGACACGGCGGGGCTTTCGGCAGATCCAGAGACTTT
  E  F  P  G  D  Y  G  W  D  T  A  G  L  S  A  D  P  E  T  F>
TGCGAAGAACAGAGAGCTGGAAGTGATCCACAGCCGGTGGGCAATGCTGGGCGCTCTGGG 1500
  A  K  N  R  E  L  E  V  I  H  S  R  W  A  M  L  G  A  L  G>
TTGCGTTTTCCCGGAGCTTCTGGCGAGAAATGGAGTAAAGTTCGGTGAGGCCGTGTGGTT
  C  V  F  P  E  L  L  A  R  N  G  V  K  F  G  E  A  V  W  F>
CAAGGCGGGAGCTCAGATTTTCAGCGAGGGAGGGCTTGACTACCTGGGAAACCCTAGCTT 1620
  K  A  G  A  Q  I  F  S  E  G  G  L  D  Y  L  G  N  P  S  L>
GGTTCACGCGCAGAGCATTCTTGCAATCTGGGCATGCCAGGTCATCCTCATGGGCGCCGT
  V  H  A  Q  S  I  L  A  I  W  A  C  Q  V  I  L  M  G  A  V>
GGAGGGCTACCGTATTGCGGGCGGTCCTCTGGGAGAGGTGACCGACCCCATCTACCCCGG 1740
  E  G  Y  R  I  A  G  G  P  L  G  E  V  T  D  P  I  Y  P  G>
GGGCAGCTTCGACCCGCTGGGACTTGCGGACGACCCCGACGCTTTCGCAGAGCTGAAGGT
  G  S  F  D  P  L  G  L  A  D  D  P  D  A  F  A  E  L  K  V>
GAAGGAGATCAAGAACGGGCGGTTGGCCATGTTTTCCATGTTCGGATTCTTCGTGCAGGC 1860
  K  E  I  K  N  G  R  L  A  M  F  S  M  F  G  F  F  V  Q  A>
CATCGTCACCGGGAAAGGACCCATCGAGAACCTCGCTGATCACCTGGCCGACCCCGTTAG
  I  V  T  G  K  G  P  I  E  N  L  A  D  H  L  A  D  P  V  S>
CAACAACGCCTGGGCCTATGCCACCAACTTCGTCCCCGGCAAATGAAATCCGGCACTAGC 1980
  N  N  A  W  A  Y  A  T  N  F  V  P  G  K  *
ACAAGTATGATGCTCTGTGATCATCTGTTCAGCCGACGGCCGGCCTGGCTTATCTCTTGT
CTATTCTGTTGTCCTTCATGTCTATTAATGCAACGAATATTTTGTATCAGATCCTTGCTA 2100
ATTTGGTTTCAGTCTTCGTTTCGAATGACTCCATTCACTTACAATTATCTTTTAAAATTA
AATCATGACAAGTGACCAATGTTGTATGTATGGTAGTTCAGTTGTTTGTTAAATTTCCTA 2220
AATTTTTGAAGGTTTTGACAGTATATAACATTGTATTGTGAACTTTTGGAACAGTGTGTG
TTACAGTTAAGGTATCCAATTTATTGGGTGAACACTTTATTTATTTTGGGAAAGAAGGAA 2340
AAGTCTCAAAAGAGTGATAAGAAGAGGTTTCAAATTAGGAANAAATATATAATTGGATAA
AAGAAGTAATATTGCAAGAAATGAAGTGGTTGGAAAGGGATAAATAATAATACTAATTTT 2460
TAACTATCATTTTTAATTATATAATTATTTTAATTATTAGAAATTAATAAATAATAATAA
TTCTCATTTTTCTTTTAATTATATAATTGTTTTTAGTTTAACATTTTTTTAATG       2574
```

Figure 3.2. The nucleotide sequence, with aligned amino acid sequence below, for the lodgepole pine *cab* gene (PNcAB2). Initiation codon is underlined and the putative TATA and CAAT boxes are in bold.

(Matsuoka et al., 1987), rice (Matsuoka 1990), barley (unpublished, Genbank) and *Lemna*19 a Type II *cab* gene (Karlin-Neumann et al. 1985) all of which are monocots. These are followed by soybean (Walling et al., 1988), cucumber (Greenland et al., 1987) and assorted other angiosperm species (Table 3.1). This suggests that the lodgepole pine *cab* gene, PNcAB2, is most closely related to other Type I *cab* genes from woody species and to monocots and less similar to dicots. This may not be surprising since it has been shown that conifers recognize monocot but not dicot promoters (Good et al., 1990). Further, the fact that PNcAB2 is more similar to a duckweed Type II *cab* than to other Type I *cab* genes like those from *Arabidopsis*, wheat (Lamppa et al., 1985) or tobacco (Castresana et al., 1987) is interesting. Even more puzzling is the greater distance between PNcAB2 and a Type II *cab* from *Pinus thunbergii*. This means that, based on sequence comparison, the duckweed Type II *cab* is more similar to a Type one *cab* gene from lodgepole pine than the same lodgepole pine is to a *cab* Type II gene from a related pine species. Most sequence variation occurs in the first 60 nucleotides of the 5' end of the mature peptide. There is 97% sequence similarity between PNcAB2 from lodgepole pine and pINEab13 from scots pine. Another *cab* gene (pINEab9) from scots pine shares 86% of its bases in the coding region; however, for this clone there is little or no homology in the untranslated regions with lodgepole pine.

Table 3.1. Comparison of the 705 bp *cab* coding region from the lodgepole pine clone PNcAB2 against selected other *cab* genes. * indicates comparison done on less than 705 nucleotides.

cab Clone	Gene Type	% Similarity to PNcAB2
Scots Pine 1B	I	97.2
Scots Pine 1A	I	87.1
Pear	I	87.0
Lemna 30	I	80.9
Maize	I	81.0
Rice	I	81.6*
Barley	I	81.0*
Lemna 19	II	80.4*
Soybean Cab4	I	77.9
Tobacco Cab7	I	77.0
Tomato CabA	I	76.9
Cucumber CabA	I	78.0
Petunia Cab25	I	77.0
Wheat	I	76.9*
Arabidopsis AB180	I	75.9
Pea Cab 8	II	76.0
Mustard	I	77.1*
Fern Cab3	II	76.9*
Pea AB80	I	75.3
Pinus thunbergii	II	76.6*
Chlamydomonas rein.	II	72.7*
Petunia Cab 37	II	73.2*
Tomato Cab4	II	72.8*
Apple	I	75.1*

Transit Peptide

Based on sequence comparison to other published *cab* transit peptides, the putative transit peptide for PNcAB2 is 40 amino acids long and this is consistent with the reported peptide size for Type I *cab* genes in scots pine but longer than the 37 amino acids reported for a Type II *cab* in *Pinus thunbergii*. This is also longer than most of the known transit peptides for Type I genes in angiosperms which range in size from 31 to 39 amino acids long. Although there is little supporting evidence a general conclusion regarding a longer transit peptide which is conifer specific for Type I *cab* genes seems unlikely. Recently it has been determined that the Japanese pear tree has a 44 residue transit peptide that differs by only a single amino acid from that reported for the larger *cab* clone in scots pine. If this pear *cab* sequence is accurate it suggests two things. First, this conserved transit sequence must be important for woody plants because it has arisen twice in the two major plant lineages, gymnosperms and angiosperms, which have been separated for more than 230 million years. Second this conserved sequence is completely different from the more typical transit peptide sequence of the Type II gene found in *P. thunbergii*. Jansson and Gustaffson (1990) point out the scots pine Type I *cab* transit peptide, and by their strong similarity, the lodgepole pine and the Japanese pear tree, are not strongly homologous with either of the reported Type I or Type II consensus transit peptides sequences of other angiosperms. The woody plants seem to be a compromise between the two angiosperm types. Unfortuately, the very close similarity between scots pine and pear, and the fact that the pear sequence is unpublished raises questions about the accuracy of the sequence. The pear *cab* gene was sequenced

by the same group that has characterized the cDNA *cab* genes from *Pinus thunbergii*, maize and rice. Perhaps it is simply contamination from the *P. thunbergii* library or perhaps a scots pine clone was used to probe the pear library resulting in some contamination. A further complication is the fact that the only other woody angiosperm *cab* sequence, from apple (Chen et al. 1990) is even more unusual and widely diverged from both Type I and II *cab* genes from angiosperms. Therefore it is difficult to attach much significance to these two sequences.

Untranslated Regions

There is no evidence of a conserved polyadenylation site or of the conserved GT motif in a logical location in PNcAB2. This is also true for the lodgepole pine *actin* gene (Kenny et al., 1988). As well, neither the Type I *cab* in scots pine or the six *rbc*S genes in larch exhibit traditional polyadenylation sites. It may be that light regulated genes in conifers utilitize a different polyadenylation signal because the conserved polyadenylation site (AAUAAA) is found in some conifers. Of possible interest as a polyadenylation sequence is the occurrence of the sequence TATGT 226 nt downstream of the termination codon in PNcAB2. This sequence also occurs in the *rbc*S genomic clone in larch 247 nt downstream of the termination codon. This conserved sequence has been found in all yeast genes studied and is located approximately 30 to 40 nt downstream from the polyadenylation site (Winnacker, 1987).

There is a five base sequence in both PNcAB2 (lodgepole pine) and pINEab13 (scots pine) which is identical to the region of the transcription start site for the light

regulated ferredoxin gene from pea (Elliot et al., 1989) and "homologous" to similar regions found in tobacco *cab*E, tomato *cab*3C, and *Chlamydomonas reinhardtii cab*II-1 (Fig. 3.3). For the *cab* genes this motif occurs around -40 to -60 nucleotides upstream to the translation start site. Unfortunately because this motif appears in the putative transcribed sequence of the cDNA from pINEab13 it is impossible for it to function as a valid transcriptional start site in lodgepole pine.

The strong sequence similarity observed between lodgepole pine and scots pine (pINEab13) extends into the untranslated 5' and 3' regions of the two genes (Fig. 3.4) suggesting that either these copies from different species are located at the same locus or that these conserved sequences play a regulatory role at the developmental or tissue specific level and that these two copies perform functions different from the other *cab* copy (pINEab9). This second *cab* copy (pINEab9) from scots pine shares 86% of its bases in the coding region; however, there is little or no homology in the untranslated regions with lodgepole pine suggesting that pINEab9 is a different locus. The 5' UTR from PNcAB2 and pINEab13 are 96% similar over 52 bases. In contrast the other *cab* copy from scots pine shares little similarity with these two except for a short four nucleotide stretch (Fig. 3.4). As well there is a strong region of homology in the first 12 bases after initiation of translation of pINEab9 and the nucleotides just preceeding the translation start codon in PNcAB2 and pINEab13.

The strong sequence similarity between PNcAB2 and pINEab13 drops to 80% in the 3' UTR and once again there is little homology with the other *cab* copy from scots pine. Nevertheless, several short stretches appear to be conserved (Fig. 3.4). The role

	-44	
Lodgepole pine	AACTGGT**ACAAGCT**AAA...	
	-44	
Scots pine pINEab13	CCGT**ACAAGCT**AAA...	
	-42	
Chlamy. reinhardtii	TTCGTTA**ACTTGCT**CAG...	
	-59	
Tomato cab3C	TGAAAAA**TCAAGCT**TCA...	
	-60	
Tobacco cabE	ATGAAAC**TCAAGCCT**CA...	
	-96	
Pea ferredoxin	TATTTCC**ACAAGCT**CCT...	

Figure 3.3. Comparison of possible initiation start sites in six *cab* genes and one ferredoxin gene. Numbers indicate location of motif in relation to the translation start site. The conserved motif is in bold letters. In the pea ferredoxin gene and the *cab* gene from *Chlamydomonas reinhardtii*, primer extension studies have identified the first A of this conserved motif as the initiation start site.

5' UPSTREAM UNTRANSLATED REGION

```
PNcAB2    CTGGTACAAGCTAAACTAATCTTCTGTTTGTGCGCCTACTGCTGCAACAATG
pINEab13  *C ***************************************************
pINEab9   TGA*AT*GGCGTAAAGC**G*AAGGAGAGAGATAGG ATGGCTACAACAATG
```

3' DOWNSTREAM UNTRANSLATED REGION

```
PNcAB2    TGAAATCCGGCACTAGCACAAGTATGATGCTCTGTGATCATCTGTTCAGCCGACGG
pINEab13  *********C*********************************T**G*****G*C*
pINEab9   TAG**GATTAACAC*A*CTCAA*CA*CA**A*A*A*CC*CATACGCTTTG*CTTGC

PNcAB2        CCGGCCTGGCTTATCTCTTGTCTATTCTGTT GTCCTTCATGTCTATTAATG
pINEab13  GGGG**************T*****************T*******C*C*G*T******
pINEab9   CC**AT**GCTCTGGCTT****CC*A*TC***A**CCCTT**TC*TCT***CC*CT

PNcAB2    CAACGAACGAATATTTTGTATCAGATCCTTGCTAATTTGGTTTCAGTCTTCGTTTC
pINEab13  *******     ***       ***********T**
pINEab9   TCCTTCCTTCC*TGAC**CG*TGT*AATC*AA*TG*A*TTC**TGCAAC*T*A*CT

PNcAB2    GAATGACTCCATTCACTTACAATTATCTTTTAAAATTAAATCATGACG
pINEab9   ACTGAGA*TTGCATTTG*TTTC**C*A
```

Figure 3.4. Sequence comparison of the 5' and 3' untranslated flanking regions of *cab* Type I PNcAB2 (lodgepole pine) and *cab* Type I pINEab13 and pINEab9 (Scots pine). The putative initiation and termination codons are in bold letters. The underlined sequences represent regions of similarity. Asteriks indicate identical nucleotides.

of these sequences is unknown; however, it is striking that such strong similarity is found within such divergent flanking regions of the same gene.

CpG Richness

Jansson and Gustafsson (1990) observed that the *cab* copies from scots pine exhibited a preference for C or G in the third position when compared to dicots but occurred less frequently than in monocots. In PNcAB2, 73% of the codons in the mature polypeptide have either C or G at the third position. This places lodgepole pine *cab* in between the two scots pine *cab* genes which were 75 and 70%. This high usage of C+G in the third position of the codon may not be common in pines. Kenny et al., (1988) found that for the carboxy portion of the lodgepole pine *actin* gene there was a slight bias (39%) against codons ending in C or G.

The coding region of the mature protein is slightly (60%) G+C rich. This is in strong contrast to the surrounding noncoding regions which are A+T rich (Fig. 3.5). It has been suggested that this feature could be correlated with genes that are highly expressed (Brinkmann et al., 1987). Also, Castresana et al., (1988) have demonstrated that in tobacco the *cab* light response elements are located in A/T rich regions upstream of the coding region. In the case of PNcAB2 the *cab* gene represents a G+C rich island surrounded by A+T rich flanking regions. This observation suggests that these A/T rich sequences may contain important regulatory information concerning gene regulation and expression. It has also been determined that lodgepole pine is generally G+C poor, with a G+C content of 37% over the genome (Miksche and Hotta, 1973).

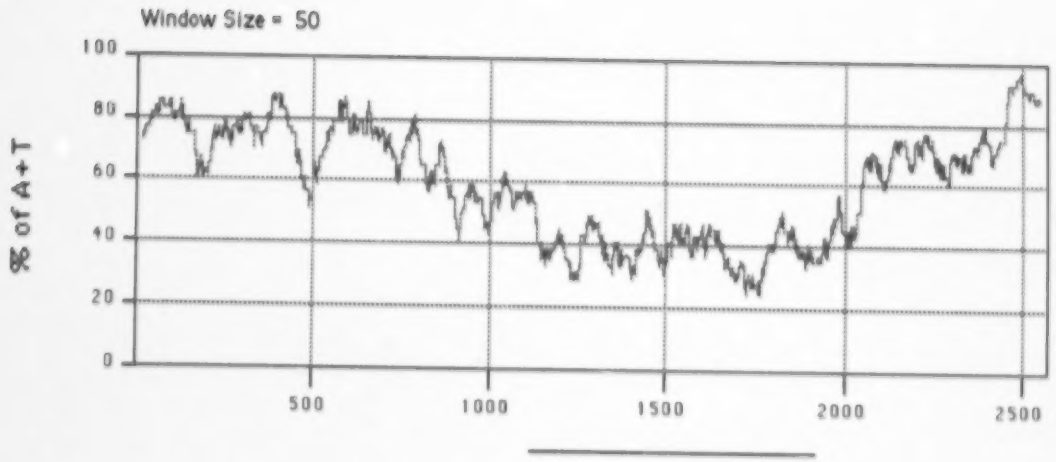


Figure 3.5. A/T composition profile of the lodgepole pine *cab* gene, PNcAB2. The coding region is underlined.

CpG Dinucleotides

Lodgepole pine exhibits a slight avoidance of possible methylation sites (CpC or CpG). In scots pine, pINEab9 showed a preference (XCG/XCC=1.41) for CpG whereas pINEab13 displayed a weak preference (XCG/XCC=0.83) for CpG. In contrast lodgepole pine (XCG/XCC=0.68) had a preference level between pINEab13 and monocots (XCG/XCC=0.50).

Promoter Region

CAB promoter regions have been well characterized in several species leading to an understanding of the regions involved in light regulation, tissue and organ specific expression and circadian patterns of *cab* expression. Comparison of the promoter region from lodgepole pine against those of pea (Simpson et al., 1985; Fluhr et al., 1986), tobacco (Castresana et al., 1988), *Arabidopsis* (Mitra et al., 1989; Ha and An, 1988) and wheat (Lamppa et al., 1985) reveal very few sequence similarities. A putative TATA and CAAT box are located 108 nt and 170 nt from the translation start codon respectively. The region between these two recognition sites has been shown to include two conserved motifs in the Type I promoters of angiosperm *cab* genes (Castresana et al., 1987). The GATA motif, which occurs one to three times in some species has been shown to be the interaction site of the DNA-protein binding factor GA-1 (Schindler and Cashmore, 1990). Mutations within the conserved GATA element have been shown to reduce expression 5-fold (Gidoni et al., 1989). In lodgepole pine there is a single perfect and two imperfect

matches to the GATA motif found between the CAAT and TATA boxes (Fig. 3.6). There are eight other copies of the GATA element upstream of the CAAT box; however, it is not known whether they are significant. Also of interest is a short conserved sequence found in both *cab* from lodgepole pine and *rbs*S from larch in approximately the same location 27 nt downstream of the CAAT box (Fig. 3.6). Whether this is a conifer specific element for light regulated genes is difficult to determine. Nevertheless, it may be of some general significance because similar sequences can be found in *Arabidopsis* cab1 (Ha and An, 1988), petunia cab22r (Dunsmuir, 1985) and soybean cab4 (Demmin et al., 1989) all located immediately downstream of the CAAT box. Of all the described areas of regulatory importance only the region 5' to the CAAT box in tobacco CAB-E shows any homology to PNcAB2.

DISCUSSION

I have constructed a genomic library for the purpose of examining the *cab* gene(s) in lodgepole pine. A homologous probe containing at least part of the *cab* coding region was used to screen the library and select a single clone which contained the complete coding region. The *cab* gene and its surrounding regions were sequenced and compared to published *cab* sequences. This copy from lodgepole pine is a Type I gene lacking an intron. It is very closely related to other *cab* genes, in particular a Type I *cab* gene from scots pine (pINE13). The two pine *cab* copies differ at 19 positions out of 705 and are 97.3% similar. These two copies are more similar than is this scots pine *cab* gene to

```
TOBACCO CAB-E  CCAATTGAAAAGTAGATATAGATACTCAAGGATAAGGGTCTTTGGGCC
LEMNA 30       CCAATGGCGTGCGGCCAGTAGATATCGGTGGATAATGATGATCCCATC
LODGEPOLE PINE GCAATTGTGGGTGTGATTGGCCGAGGAGATTCGGAAACGATAGGAGAT
LARCH RBSC     CCAATTTCTCTGGTCTGGCCCAACCTTAGAGATTGGGGCTTATATCTA
SOYBEAN CAB4   CCAATAGGAATAAAAATATAGATCTCATAAGATATTAAGATAAGATAA
PETUNIA CAB22R CCAATTAGGAAATAGATAGTGATTATCAAGGATAAGGACTTAGGGTCT
ARABODOPSIS    CCAATGAATGAACAGATAAAGATTACTTCAGATATAACAAACGTTACA
```

```
TOBACCO CAB-E  TGTAAAGCCATTTATA
LEMNA 30       CTCGTCTCCTTCCTCTTCGGTATTA
LODGEPOLE PINE GAAATTATTTGGTCTGCGGTTGCTCTATATATT
LARCH rbcS
SOYBEAN CAB4   CATTTTCTATTTGAAAGCTATATATA
PETUNIA 22R    TTCGAGTCATTTAAATAAA
ARABIDOPSIS    ATATCCCTATATA
```

Figure 3.6. Comparison of promoter region between the CAAT and TATA box. The CAAT and TATA boxes as well as the conserved GATA motifs are in bold. The larch *rbcS* sequence is absent from the bottom paragraph because the distance between the CAAT and TATA boxes is much shorter than similar sequences for the *cab* genes. Conserved sequences between species are underlined. All *cab* genes are Type I except for the larch *rbcS* sequence.

another Tvpe I gene (pINE9) in the same species. They differ at 97 positions out of a possible 705 or are 13.8% diverged. The similarity between lodgepole pine and scots pine extended into the 5' and to a lesser extent the 3' untranslated regions. This high degree of similarity between species has not been observed before, most likely because of the lack of sequence data between closely related species. Most studies have concentrated on within-species characterization. Nevertheless examination of the only other congeneric species, *Chlamydomonas reinhardtii* (Imbault et al., 1988) and *C. moewusii* (Larouche et al., 1991) indicates that the published *cab* genes from these species are only 86% similar.

It is disappointing that more similarities were not found between the 1 kb region upstream of the coding region in lodgepole pine and the known 5' regulatory regions outlined in tobacco (Schindler and Cashmore, 1990; Castresana et al., 1988), wheat (Fejes et al., 1990; Nagy et al., 1988), pea (Fluhr et al., 1986; Simpson et al., 1986a; Simpson et al., 1986b; Simpson et al., 1985) and *Arabidopsis* (Mitra et al, 1989; Ha and An, 1988). Nevertheless, the occurrence of an A/T rich upstream region was also observed in tobacco (Castresana et al., 1987). The identified regulatory elements in the tobacco promoter region were found to contain 70-80% A/T which is a level similar to that found in lodgepole pine. Unfortunately, correct identification of light regulatory elements (LRE) and, negative and positive regulatory elements (NRE; PRE) in pines may only be achieved through deletion analysis of the promoter regions. The absence of any homology to the tobacco LRE is not surprising for two reasons. First, evidence is accumulating that conifers are not strongly regulated by light (Alosi et al. 1990;

Yamamoto et al. 1991). Perhaps lodgepole pine does not require a LRE. Second, comparison of the regulatory region from wheat does not match any similar elements in pea or tobacco. As well, pea is not similar to tobacco therefore the regulation of *cab* genes may be species specific. I believe that the elucidation of the regulatory control of *cab* genes may provide some insight into the phenomenon of shade intolerance in pines. It is conceivable that lodgepole pine has evolved a structural organization, either within the promoter region or within the thylakoid membranes, that makes light capture and transfer less efficient than shade tolerant conifers, such as *Picea* or *Tsuga*. Further isolation and characterization of *cab* genes in other conifers is desirable and may be informative.

LITERATURE CITED

Alosi, C.M., D.B. Neale and C.S. Kinlaw. 1990. Expression of *cab* genes in douglas-fir is not strongly regulated by light. Plant Physiol. 93:829-832.

Barnes, W.M., M. Bevan and P.H. Son. 1983. Kilo-sequencing: creation of an ordered nest of asymetric deletions across a large target sequence carried on phage M13. Methods Enzymol. 101:98-122.

Batschauer A., B. Ehmann and E. Schaefer. 1991. Cloning and characterization of a chalone synthase gene from mustard and its light-dependent expression. Plant Mol. Biol. 16:175-185.

Brinkmann, H., P. Martinez, F. Quigley, W. Martin, and R. Cerff. 1987. Endosymbiotic origin and codon bias of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. J. Mol. Evol. 26:320-328.

Buetow, D.E., H. Chen, G. Erdos and L.S. Yi. 1988. Regulation and expression of the multigene family coding light-harvesting chlorophyll a/b-binding proteins of photosystem II. Photosyn. Res. 18:61-97.

Butler, P.J.G. and W. Kuhlbrandt. 1988. Determination of the aggregate size in detergent solutions of the light-harvesting chlorophyll a/b-protein complex from chlorophlast membranes. Proc. Nat. Acad. Sci. 85:3797-3801.

Cashmore, A. 1984. Structure and expression of a pea nuclear gene encoding a chlorophyll a/b-binding polypeptide. PNAS. 81:2960-2964.

Castresana, C., I. Garcia-Luque, E. Alonso, V.S. Malik and A.R. Cashmore. 1988. Both positive and negative regulatory elements mediate expression of a photoregulated CAB gene from *Nicotiana plumbaginifolia*. EMBO J. 7:1929-1936.

Castresana, C., R. Staneloni, V.S. Malik and A.R. Cashmore. 1987. Molecular characterization of two clusters of genes encoding the Type I CAB polypeptides of PSII in *Nicotiana plumbaginifolia*. Plant Mol. Biol. 10:117-126.

Chen, H., S.S. Korban, and D.E. Buetow. 1990. Nucleotide sequence of an apple nuclear gene encoding a light-harvesting chlorophylla/b binding polypeptide of photosystem II. Nucl. Acids Res. 18:679.

Demmin, D.S., E.J. Stockinger, Y.C. Chang and L.L. Walling. 1989. Phylogenetic relationships between the chlorophyll a/b binding protein (CAB) multigene family: an intra- and interspecies study. J. Mol. Evol. 29:266-279.

Dunsmuir, P. 1985. The petunia chlorophyll a/b binding protein genes: a comparison of Cab genes from different gene families. Nucl. Acids Res. 13:2503-2518.

Elliott, R.C., T.J. Pedersen, B. Fristensky, M.J. White, L.F. Dickey and W.F. Thompson. 1989. Characterization of a single copy gene encoding ferredoxin I from pea. Plant Cell 1:681-690.

Fejes, E., A. Pay, I. Kanevsky, M. Szell, E. Adam, S. Kay and F. Nagy. 1990. A 268 bp upstream sequence mediates the circadian clock-regulated transcription of the wheat Cab-1 gene in transgenic plants. Plant Mol. Biol. 15:921-932.

Fluhr, R., C. Kuhlemeier, F. Nagy and N-H. Chua. 1986. Organ-specific and light-induced expression of plant genes. Science 232:1106-1112.

Gidoni, D., P. Brosio, D. Bond-Nutter, J. Bedbrook and P. Dunsmuir. 1989. Novel cis-acting elements in petunia *cab* gene promoters. Mol. Gen. Genet. 215:337-344.

Good, A.G., F. Bekkaoui, G. Polate, D.I. Dunstan and W.L. Crosby. 1990. Anaerobic induction in conifers: expression of endogenous and chimeric anaerobically-induced genes. Physiolog. Plantarum 78:441-446.

Green, B.R., E. Pichersky and K. Kloppstech. 1991. Chlorophyll a/b- binding proteins: an extended family. TIBS 16:181-186.

Greenland, A.J., M.V. Thomas and R.M. Walden. 1987. Expression of two nuclear genes encoding chloroplast proteins during early development of cucumber seedlings. Planta 170:99-110.

Ha, S-B. and G. An. 1988. Identification of upstream regulatory elements involved in the developmental expression of the *Arabidopsis thaliana* cab1 gene. Proc. Nat. Acad. Sci. 85:8017-8021.

Henikoff, S. 1987. Unidirectional digestion with exonuclease III in DNA sequence analysis. Methods Enzymol. 155:156-165.

Hohn, B. 1979. In vitro packaging of λ and cosmid DNA. Methods Enzymol. 68:229.

Hutchison, K.W., P.D. Harvie, P.B. Singer, A.F. Brummer and M.S. Greenwood. 1990. Nucleotide sequence of the small subunit of ribulose-1,5-bisphosphate carboxylase from the conifer *Larix laricina*. Plant Mol. Biol. 14:281-284.

Imbault, P., C. Wittemer, U. Johanningmeier, J.D. Jacobs and S.T. Howell. 1988. Structure of the *Chlamydomonas reinhardtii cab*II-1 gene encoding a chlorophyll a/b-binding protein. Gene 73:397-407.

Jansson, S., and P. Gustafsson. 1990. Type I and Type II genes for the chlorophyll a/b-binding protein in the gymnosperm *Pinus sylvestris* (Scots pine):cDNA cloning and sequence analysis. Plant Mol. Biol. 14:287-296.

Karlin-Neumann, G.A., B.D. Kohorn, J.P. Thornber and E.M. Tobin. 1985. A chlorophyll a/b-protein encoded by a gene containing an intron with characteristics of a transposable element. J. Mol. Appl. Genet. 3:45-61.

Kenny, J.R., B.P. Dancik, L.Z. Florence and F.E. Nargang. 1988. Nucleotide sequence of the carboxy-terminal portion of the lodgepole pine actin gene. Can. J. For. Res. 18:1595-1602.

Kohorn, B., E. Harel P.R. Chitnis, J.P. Thornber and E.M. Tobin. 1986. Functional and mutational analysis of the light-harvesting chlorophyll a/b protein of thylakoid membranes. J. Cell Biol. 102:972-981.

Kuhlbrandt, W. 1984. Three-dimensional structure of the light-harvesting chlorophyll a/b-protein complex. Nature. 307:478-480.

Kuhlbrandt, W., and D.N. Wang. 1991. Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. Nature 350:130-134.

Lamppa, G.K., G. Morelli and N-H Chua. 1985. Sturcture and developmental regulation of a wheat gene encoding the major chlorophyll a/b-binding polypeptide. Mol. Cell. Biol. 5:1370-1378.

Larouche, L., C. Tremblay, C.Simard, and G. Bellemare. 1991. Characterization of a cDNA encoding a PSII-associated chlorophyll a/b-binding protein (CAB) from *Chlamydomonas moewusii* fitting into neither type I nor type II. Curr. Genet. 19:285-288.

Leutwiler, L.S., E.M. Meyerowitz and E.M. Tobin. 1986. Structure and expression of three light-harvesting chlorophyll a/b-binding protein genes in *Arabidopsis thaliana*. Nucl. Acids Res. 14:4051-4064.

Long, Z., S. Wang and N. Nelson. 1989. Cloning and nucleotide sequence analysis of genes coding for the major chlorophyll-binding protein of the moss *Physcomitrella patens* and the halotolerant alga *Dunaliella salina*. Gene 76:299-312.

Mason, J.G. 1989. Nucleotide sequence of a cDNA encoding the light-harvesting chlorophyll a/b binding proteint from spinach. Nucl. Acids Res. 17:5387.

Matsuoka, M. 1990. Classification and characterization of cDNA that encodes the light-harvesting chlorophyll a/b binding protein of photosystem II from rice. Plant Cell Physiol. 31:519-526.

Matsuoka, M., Y. Kano-Murakami and N. Yamamoto. 1987. Nucleotide sequence of cDNA encoding the light-harvesting chlorophyll a/b binding protein from maize. Nucl. Acids Res. 15:6302.

Mead, D.A., E. Szczesna-Skorupa and B. Kemper. 1986. Single-stranded DNA 'blue' T7 promoter plasmid: a versatile tandem promoter system for cloning and protein engineering. Prot. Engineer. 1:67-74.

Miksche, J.P. and Y. Hotta. 1973. DNA base compositions and repetitious DNA in several conifers. Chromosoma 41:29-36.

Mitra, A., H.K. Choi and G. An. 1989. Structural and functional analyses of *Arabidopsis thaliana* chlorophyll a/b-binding protein (*cab*) promoters. Plant Mol. Biol. 12:169-179.

Nagy, F., S.A. Kay and N-H. Chua. 1988. A circadian clock regulates transcription of the wheat *Cab*-1 gene. Genes and Dev. 2:376-382.

Pichersky, E., R. Bertanzky, S.D. Tanksley, R.W. Breidenbach, A.P. Kausch and A.R. Cashmore. 1985. Molecular characterization and genetic mapping of two clusters of genes encoding chlorophyll a/b-binding proteins in *Lycopersicon esculentum* (tomato). Gene 40:247-258.

Pichersky, E., N.E. Hoffman, V.S. Malik, R. Bernatzky, S.D. Tanksley, L. Szabo and A.R. Cashmore. 1987a. The tomato *Cab*-4 and *Cab*-5 genes encode a second type of CAB polypeptides localized in Photosystem II. Plant Mol. Biol. 9:109-120.

Pichersky, E., N.E. Hoffman, R. Bernatsky, B. Piechulla, S.D. Tanksley and A.R. Cashmore. 1987b. Molecular characterization and genetic mapping of DNA sequences encoding the Type I chlorophyll a/b-binding polypeptide of photosystem I in *Lycopersicon esculentum* (tomato). Plant Mol. Biol. 9:205-216.

Pichersky, E., S.D. Tanksley, B. Piechulla, M. Stayton and P. Dunsmuir. 1988. Nucleotide sequence and chromosomal location of *Cab*-7, the tomato gene encoding the Type II chlorophyll a/b-binding polypeptide of Photosystem I. Plant Mol. Biol. 11:69-71.

Pichersky, E., T.G. Brock, D. Nguyen, N.E. Hoffman, B. Piechulla, S.D. Tanksley and B.R. Green. 1989. A new member of the CAB gene family: structure, expression and chromosomal location of *Cab*-8, the tomato gene encoding the Type III chlorophyll a/b-binding polypeptide of photosystem I. Plant Mol. Biol. 12:257-270.

Pichersky, E., D. Soltis and P. Soltis. 1990. Defective chlorophyll a/b-binding protein genes in the genome of a homosporous fern. Proc. Nat. Acad. Sci. 87:195-199.

Rigby,P.W.J., M. Dieckmann, C. Rhodes, and P. Berg. 1977. Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. 113:237.

Rosenberg, S.M. 1985. *Eco*K restriction during *in vitro* packaging of coliphage lambda DNA. Gene 39:313.

Rosenberg, S.M. 1987. Improved in vitro packaging of lambda DNA. Methods Enzymol. 153:95.

Sambrook, J., E. Fritsch and T. Maniatis. 1989. *Molecular Cloning*. Cold Spring Harbour Laboratory Press.

Schindler, U. and A.R. Cashmore. 1990. Photoregulated gene expression may involve ubiquitous DNA binding proteins. EMBO J. 9:3415-3427.

Simpson, J., J. Schell, M. Van Montagu and L. Herrera-Estrella. 1986a. Light-inducible and tissue-specific pea lhcp gene expression involves an upstream element combining enhancer- and silencer-like properties. Nature 323:551-554.

Simpson, J., M.P. Timko, A.R. Cashmore, J. Schell, M. Van Montagu and L. Herrera-Estrella. 1985. Light-inducible and tissue-specific expression of a chimaeric gene under control of the 5'-flanking sequence of a pea chlorophyll a/b-binding protein gene. EMBO J. 4:2723-2729.

Simpson, J., M. Van Montagu and L. Herrera-Estrella. 1986b. Photosynthesis-associated gene families: differences in response to tissue-specific and environmental factors. Science 233:34-38.

Stayton, M.M., M. Black, J. Bedbrook and P. Dunsmuir. 1986. A novel chlorophyll a/b binding (*Cab*) protein gene from petunia which encodes the lower weight *Cab* precursor protein. Nucl. Acids Res. 14:9781-9796.

Walling, L.L., Y.C. Chang, D.S. Demmin and F.M. Holzer. 1988. Isolation, characterization and evolutionary relatedness of three members from the soybean multigene family encoding chlorophyll a/b binding proteins. NAR 16:10477-10492.

Watson, J.C., and W.F. Thompson. 1986. Purification and restriction endonuclease analysis of plant nuclear DNA. Methods Enzymol. 118:57-74.

Westneat, D.L., W.A. Noon, H.K. Reeve and C.F. Aquadro. 1988. Improved hybridization conditions for DNA 'fingerprinting' probed with M13. Nucl. Acids Res. 16:4161.

Winnacker, E.L. 1987. *From Genes to Clones*. VCH, Heidelberg, FRG.

Yamamoto, N., Y. Kano-murakami, M. Matsuoka, Y. Ohashi and Y. Tanaka. 1988a. Nucleotide sequence of a full length cDNA clone of ribulose bisphosphate carboxylase small subunit from green dark-grown pine (*Pinus thunbergii*) seedling. Nucleic Acids Res. 16:11830.

Yamamoto, N., M. Matsuoka, Y. Kano-Murakami, Y. Tanaka and Y. Ohashi. 1988b. Nucleotide sequence of a full length cDNA clone of light harvesting chlorophyll a/b binding protein gene from green dark-grown pine (*Pinus thunbergii*) seedling. Nucleic Acids Res. 16:11829.

Yamamoto, N., Y. Mukai, M. Matsuoka, Y. Kano-Murakami, Y. Tanaka, Y. Ohashi, Y. Ozeki and K. Odani. 1991. Light-independent expression of *cab* and *rbc*S genes in dark-grown pine seedlings. Plant Physiol. 95:379-383.

CHAPTER FOUR

EXAMINATION OF THE CAB TRANSIT PEPTIDES

INTRODUCTION

Many of the proteins involved in photosynthesis are encoded by nuclear genes and synthesized on cytosolic ribosomes. Amino terminal extensions on these polypeptides assist the transport and trafficking of these proteins into the chloroplast (Chua and Schmidt, 1979). These extensions, chloroplastic transit peptides (cTP), have been characterized for numerous proteins and include chlorophyll a/b binding proteins (CAB), ribulose-bisphosphate carboxylase/oxygenase (SSU), ferredoxin, plastocyanin, and early light inducible proteins (ELIPs). Based on differences between the molecular weight of the mature and the precursor polypeptides, as well as comparison of the processing site of various *rbc*S transit peptides, the *cab* transit peptide cleavage site has been located just prior to the M-R-K motif of the mature polypeptide (Karlin-Neumann et al. 1985). Amino acid sequence analysis of available cTPs have led to several general observations (Gavel and Von Heijne 1990; Keegstra et al., 1989). The common features of chloroplast transit peptides include high levels of hydroxylated serines and threonines, a richness of hydrophobic amino acids, such as alanine and valine, and a lack of acidic amino acids (Keegstra 1989, Von Heijne et al., 1989). It has also been suggested that cTPs contain three structural domains (Karlin-Neumann and Tobin 1986; Von Heijne et al., 1989) and a loosely defined cleavage recognition site (Val/Ile)-X-(Ala/Cys)*Ala, where the asterik denotes the proposed cleavage site (Gavel and Von Heijne 1990).

In constrast, the size of cTPs range from 19 amino acids (*Chlamydomonas reinhardtii* CAB, Imbault et al., 1988) to more than 75 amino acids (petunia 5-enolpyruvyl-shikimate-3-phosphate synthase, Klee et al., 1987). Although sequence similarities exist among transit peptides of the same precursors from different plant species, few similarities are found among different precursors from the same plant species (Von Heijne et al., 1989).

The chlorophyll a/b binding protein belongs to a class of nuclear encoded genes and is synthesized in the cytosol. It is then transported into the chloroplast. Directly upstream of the coding region for the mature peptide is the genetic information for synthesis of a transit peptide which allows/assists the mature peptide to be transported across the chloroplast membrane. Together, this apoprotein is transported into the stroma where the transit peptide is cleaved by a stromal peptidase (Keegstra et al., 1989) and the mature peptide is directed into the thylakoid membrane. It is unclear how the encoded information is used by the chloroplast machinery; however, *in vivo* experiments with chloroplast and mitochondrial transit peptides suggest that the targeting sequences are very specific for their respective organelles (Boutry et al., 1987). Nevertheless, the CAB transit peptide appears to only direct the apoprotein into the stroma. If the precursor sequence is replaced by the transit peptide from SSU, CAB is imported and integrated into the thylakoid membrane with equal efficiency (Lamppa, 1988). This suggests that the transit peptide is not protein specific. It seems likely that the transit peptide is recognized by a chloroplast envelope receptor and the apoprotein is able to bind to the chloroplast membrane and pass into the stroma. The mechanism of transport and

insertion within the chloroplast for CAB is unknown; however, it is believed to be directed by an unidentified sequence, probably found within the mature protein (Keegstra, 1989; Kohorn et al., 1986).

A comparison of the transit peptide for the Type I and II LHC proteins suggest that the CAB transit peptides do not fit the general cleavage scheme of cTPs and that they lack many of the common features associated with cTPs.

MATERIALS AND METHODS

Sequences

Sequences were compiled from the literature or from the EMBL Gene Sequence Library in Heidelberg. DNA sequences were translated using MacVector 2.0 (IBI International). Alignment of sequences was done visually or based on previously published arrangements.

Hydrophilicity Analysis

Hydrophilicity profiles were calculated by MacVector. Hydrophilicity values for each amino acid were derived using the Kyte-Doolittle scale (Kyte and Doolittle, 1982). These values are assigned using a combination of water vapour transfer free energies for amino acid side chains and the preference of amino acid side chains for interior or exterior environments (Kyte and Doolittle, 1982). Although the scale was originally developed for hydrophobicity plots, MacVector simply reverses the signs and plots hydrophilicity (MacVector Manual, 1990).

Hydrophobic Moment Analysis

MacVector amphiphilicity profiles were used to detect amphiphilic regions in selected CAB transit peptides. MacVector calculates the periodicity of the protein's hydrophobic moment based on Eisenberg's normalized consensus scale (Eisenberg et al., 1984) over a window size of 10 residues. It has been determined that an amphiphilic alpha helix will exhibit a periodicity of 100 degrees. As well, a beta strand will have a periodicity of 180 degrees. In practice the periodicity may range from 160 to 180 degrees as a result of bends and twists in a sheet (MacVector Manual, 1990). Therefore, the program assumes a periodicity of 170 degrees as a compromise. Values greater than 0.4 indicate amphiphilic regions.

Graphs were prepared for only a single representative of each group to determine the amphiphilic profile.

RESULTS

Lodgepole Pine Transit Peptide

The transit peptide for the chlorophyll a/b binding protein for lodgepole pine is estimated to be 40 amino acids long (Table 4.1). The protein sequence is 97% similar to the transit peptide of scots pine, pINE13, and 72.5% homologous to another scots pine, pINE9 and to a pear *cab* transit peptide (unpublished, Kano-Murami 1991; in EMBL Library). The 40 amino acids are longer than other reported transit peptides of both

Table 4.1. Compilation of CAB transit peptides, aligned within groups. The first sequence in each group is listed by its complete sequence. Periods indicate identical amino acids. Dashes represent deletions for the purposes of alignment. Sequences are aligned around the three amino acids, indicated with bold letters, that are conserved among sequences. Amino acids underlined at the N-terminal of lodgepole and scots pine indicate a feature discussed in the text but is not part of the proposed transit peptide sequence.

TYPE I CAB TRANSIT PEPTIDES

```
SOYBEAN 2   MAAS-T--MALSSSS-LAGQAMKLAPSTPEL-GVGRVS
SOYBEAN 3   ...ASS--.....P.-...K.V..G..A..---.....
SOYBEAN4    ....-.--.......-.....I.......Q.-......
SOYBEAN 5   ....-.--.......-.....I.......Q.-......
PEA 80      ....SSSS.....PT-...KQL..N..SQ..-.AA.FT
TOMATO 1A   ...AA---.....P.-F....V..S..AS.NS.N..IT
TOMATO 1B   ...A-.--.....P.-F....V..S..AS.IS.N..IT
TOMATO 1C   ...A-.--.....P.-F....V..S..AS.IS.N..IT
TOMATO 3A   ....-.--......T-F..KTV.....SS.IT.N..IT
TOMATO 3B   ....-.--......T-F..K.V..S..SS.IS.N..IT
TOMATO 3C   ..T.-.--......T-F..K.V..S..SS.IT.N..IT
PETUNIA91R  ...AA.--.....P.-F..K.V.FS..SS.IT.N.KAT
PETUNIA13   ...A-.--.......-F..K.VNV-..SS.ITRN.K.T
PETUNIA25   ...A-.--..I....-F..K.VNVS..SSQIT.N.KAT
PETUNIA22L  ...A-.--......T-F..KVV..S..SS.IT.N.KAT
PETUNIA22R  ...T-.--.......-F..K.V..SS.SS.IT.N.K.T
ARABID140   ....-.--.....PA-F..K.V..S.AAS.VL.S...T
ARABID165   ....-.--.....PA-F..K.VN.S.AAS.VL.S...T
ARABID180   ....-.--.....PA-F..K.VN.S.AAS.VL.S...T
LEMNA30     ....---.....P.-.V.K.V....AAS.VF.E....
TOBACCO C   ....-.--.......-F..K.V..S..SS.IT.N.K.T
TOBACCO E   ....-.--T....PF---.K.V..S..SS.VT.N.K.T
MUSTARD     ....-.--.....PA-F..K.V..S.GAS.VF.T...T
WHEAT 1.6   ...T-.--.S.....-...K.V.NL..SALI-.DA..N
MAIZE       ..S.-.--.....TA-F..K.VNV-..SS-F-.EA..T
SPINACH     ..S.-.--.....P.-...K.V..G.TAS.II.E..IT
CUCUMBER                PT-.....V..S.NA..IE.NAKFT
```

TYPE II CAB TRANSIT PEPTIDES

```
PETUNIA 37      MAT-SAIQQSAFAGQTALKSQ-NELVRKIGSFGGGRAT
TOMATO 4        ...-C.......V..AVG...-..FI..V.N..E..I.
TOMATO 5        ...-C.......V..AVG...-..FI..V...E...V.
PINE TUN        ...A....S.SL....L.RP.Q....K.V.-TAQA.I.
SILENE          ...-.T.........L..P.-......V.GN-...VS
LEMNA 19        ..A-..L-S....C.-...-.RD.....V..----.FS
RICE2           ..A-..LH.TTS.L.-..-PRRDDV..RV.D-S...I.
```

WOODY PLANTS CONTAINING TYPE I CAB TRANSIT PEPTIDES

```
PNcAB2      TAATMASCGIGS-CGFAGGQISSLKPHTNQLLGVGAGVHGEARVT
pINEAB13    ....................-.A.............................
pINEAB9     M.T.........R.A...A.L..V..QN.......-.A..Q..L.
PEARCAB     M.T.........R.A..FA.L..V..QN.......-.A.....L.
```

OTHER UNDEFINED CAB TRANSIT PEPTIDES

```
APPLE           MYIPRLSVSICLHPLLIYKIAKPSKLNPISSTLIVTAGFT
FERN CAB3         MATSTAALTSTFSGQQL-KPVNELSRKV-GAGEARVQ
PHYSCO MOSS       ..AA..MHSTALA..S.V.........L.NV....T
DUNAL ALGAE          MQLSTPPEVPDMKNTYSNNECPAAEQELCSEQGGRAGKST
CHLAMY REIN          MAFALASRKALQVTCKAT
CHLAMY MOE           MAFALISNRALTVKVEAAKKAAG
```

Types I and II *cab* genes which range in length from 33-36 amino acids. Although the transit peptide from lodgepole pine is clearly similar to those of scots pine and pear, there is little evidence of amino acid sequence similarity between these woody plants and the other reported transit peptides (Table 4.1). Only one short region of homology, the sequence F-A-G exists among the various published sequences.

Lodgepole pine's transit peptide is generally hydrophobic (Fig. 4.1) with a large, 10 amino acid, hydrophobic region juxtaposed to the proposed cleavage site. It is tempting to describe this region as a transmembrane domain, perhaps functioning in the transport of the mature LHC into the chloroplast. Although this hydrophobic region occurs in the scots pine transit peptides, it is lacking in the other CAB transit peptides.

Amino Acid Frequency Analysis

It has been suggested that chloroplast transit peptides share several features. These include the hydroxylated amino acids, serine and threonine, comprising 20-35% of the total peptide residues, a higher concentration of the hydrophobic amino acids, alanine and valine, and a deficiency of acidic residues (Keegstra et al., 1989). The transit peptide from lodgepole pine does not seem to share these common features. Serine is present at 10% of the sites and threonine occurs only 5% of the time. This is less than the 20-35% expected based on cTP studies (Von Heijne et al., 1989). Nevertheless, closer examination of other CAB transit peptides from Type I and II apoproteins reveals that serine and threonine occur 11% and 8%, respectively, in Type II CAB transit peptides and 27% and 6% in Type I transit peptides. In all cases, threonine is less likely to occur in

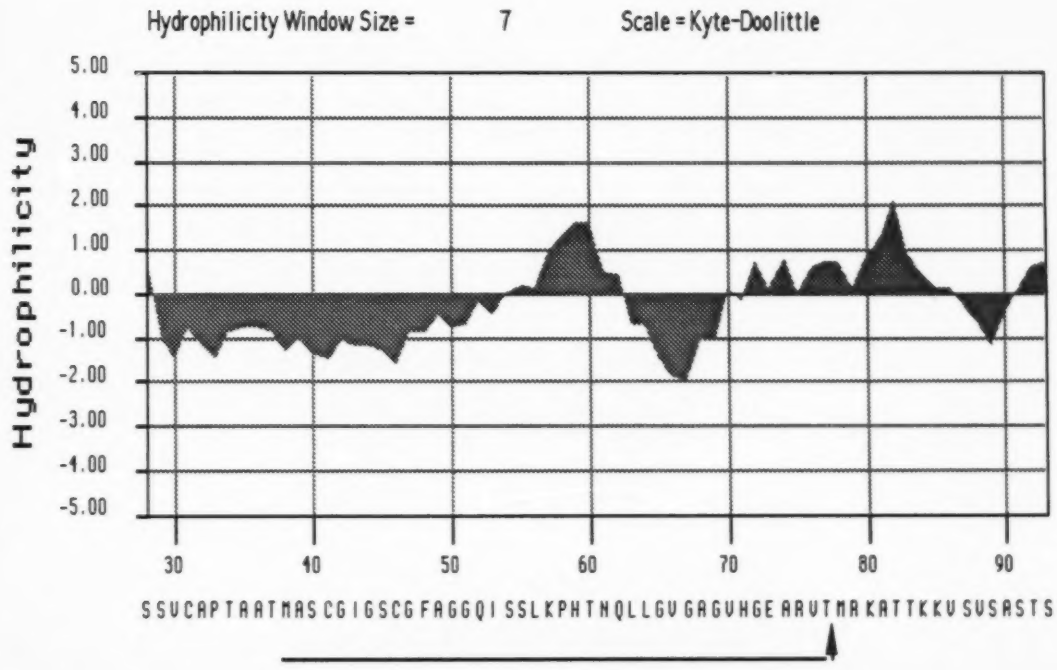

Hydrophilicity Window Size = 7
Scale = Kyte-Doolittle
Hydrophilicity
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-4.00
-5.00
30
40
50
60
70
80
90
S S V C A P T A A T M A S C G I G S C G F A G G Q I S S L K P H T N Q L L G V G A G V H G E A R V T M A K A T T K K V S V S A S T S

Figure 4.1. Hydrophilicity profile of the lodgepole pine transit peptide and surrounding region. The proposed transit peptide is underlined. The arrow indicates the proposed cleavage site.

CAB transit peptides than in chloroplast transit peptides in general. Serine is more common in Type I transit peptides but is deficient in Type II transit peptides.

Lodgepole pine does not share the richness of alanine and valine that has been reported for the other chloroplast transit peptides. Ten percent of the residues are alanines and 7 % are valines. These values are consistent across PSII CAB transit peptides with alanine occurring 15% and 11% in Type I and Type II CAB transit peptides, respectively. Valine is observed in 6 and 5% of the positions in Type I and II transit peptides, respectively.

In all cases, the CAB transit peptides are deficient in acidic amino acids. Glutamic acid occurs once in lodgepole pine, once in most Type I transit peptides and 1-2 times in Type II precursors. There are no acidic residues in scots pine pINE9 and one in pINE13. As well, there is no evidence of aspartic acid in CAB transit peptides except for the Type II CABs Lemna19 (occurs once), and rice2 (3 times), and the alga *Dunaliella salina* (once).

In contrast, glycine is more common in lodgepole pine than is expected in chloroplast transit peptides. It occurs in 22% of the positions in lodgepole pine and scots pine. It ranges from 9-11% in the other CAB transit peptides. It has been observed that the N-terminal 5-10 residues lack both positively charged amino acid as well as glycine and proline residues (Von Heijne et al., 1989). In lodgepole pine, glycine occurs three times in the first ten amino acids of the N-terminal; however, it does not occur in the N-terminal of other CAB transit peptides.

Hydrophobic Moment

Hydrophobic moment analysis of the transit peptide reveals a pronouced peak at the C-terminal end of the transit peptide (Fig. 4.2a). There is no evidence of an amphiphilic region at the N-terminal nor of an alpha helix structure at the C-terminal end. In contrast there is a strong indication that the carboxy end forms an amphiphilic beta sheet. Similar profiles were obtained for petunia 22L and rice2, representing typical Type I and Type II CAB transit peptides, respectively (Fig. 4.2B,C). In both cases the peaks were not as pronounced as for lodgepole pine, nevertheless the last 8 residues of petunia 22L and the last 5 amino acids of rice displayed this trend. Interestingly, neither lodgepole pine nor petunia 22L showed any indication of an amphiphilic helix structure in the peptide. In contrast, there is a strong peak for rice just before the amphiphilic sheet structure (Fig. 4.2C). It is unclear whether this is significant. The final CAB transit peptide that was analyzed was apple (Fig. 2.4D). Surprisingly, it does not support the general trend of a amphiphilic beta strand at the C-terminal. In fact, the profile indicates no amphiphilic structures in the C-terminal portion of that transit peptide. However, there is a suggestion of an amphiphilic beta sheet composed of the eight N-terminal amino acids. This is in direct contrast to the other CAB transit peptides and further illustrates the differences between these species. Amphiphilic Beta-strand structure has been predicted for chloroplast transit peptides and may be related to membrane surface binding and/or receptor binding on the chloroplast (Von Heijne et al., 1989).

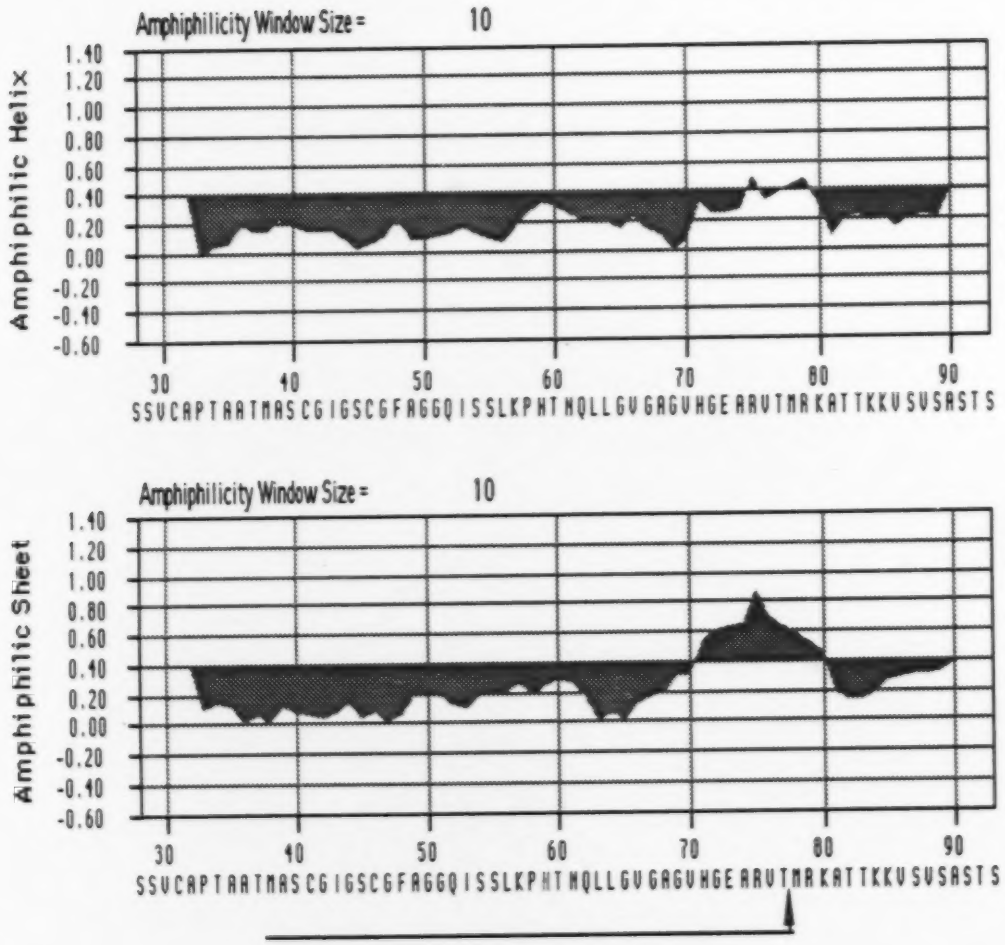


Figure 4.2A. Amphiphilic profiles of the lodgepole pine transit peptide region. Peptide sequence is underlined and the arrow indicates the proposed site of peptide cleavage.

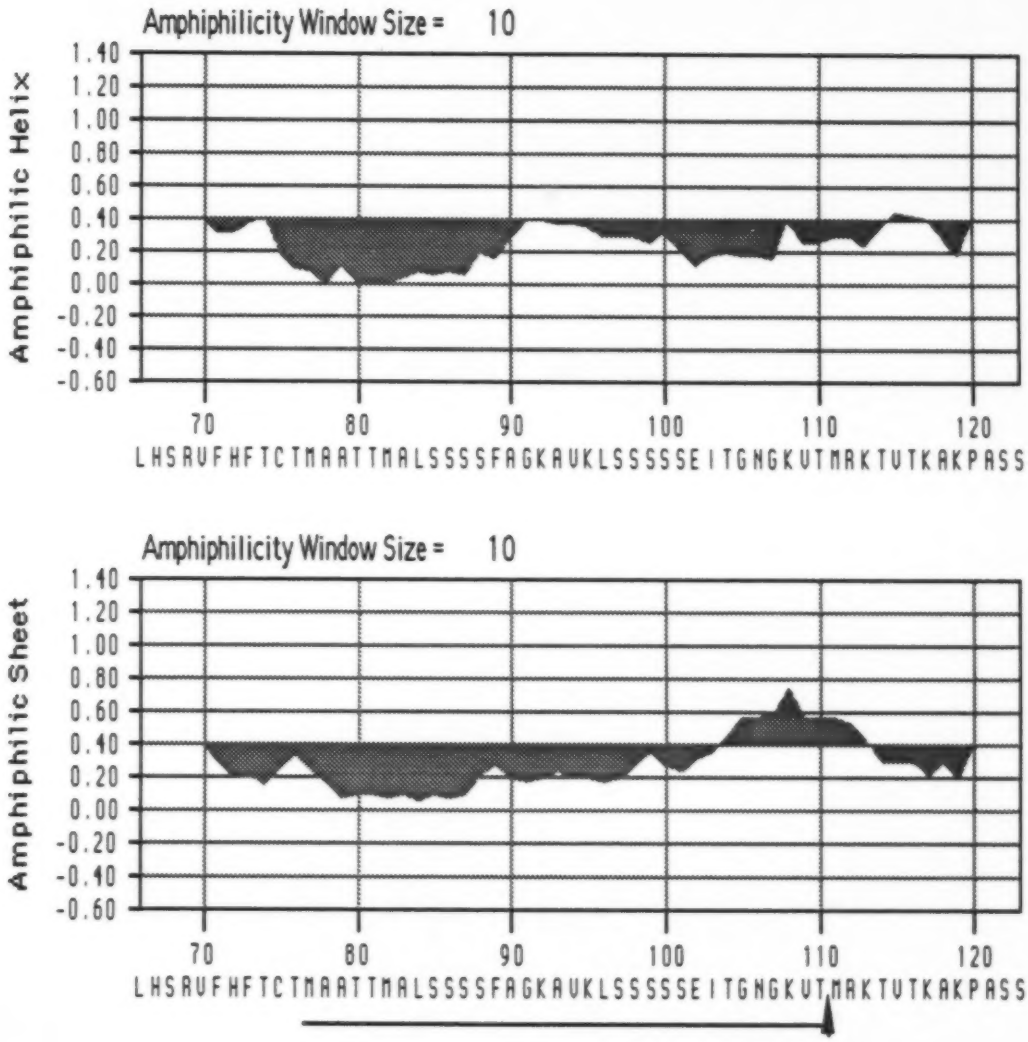


Figure 4.2B. Amphiphilic profile of the transit peptide and surrounding region of petunia 22R. The transit peptide sequence is underlined. The proposed cleavage site is indicated with an arrow.

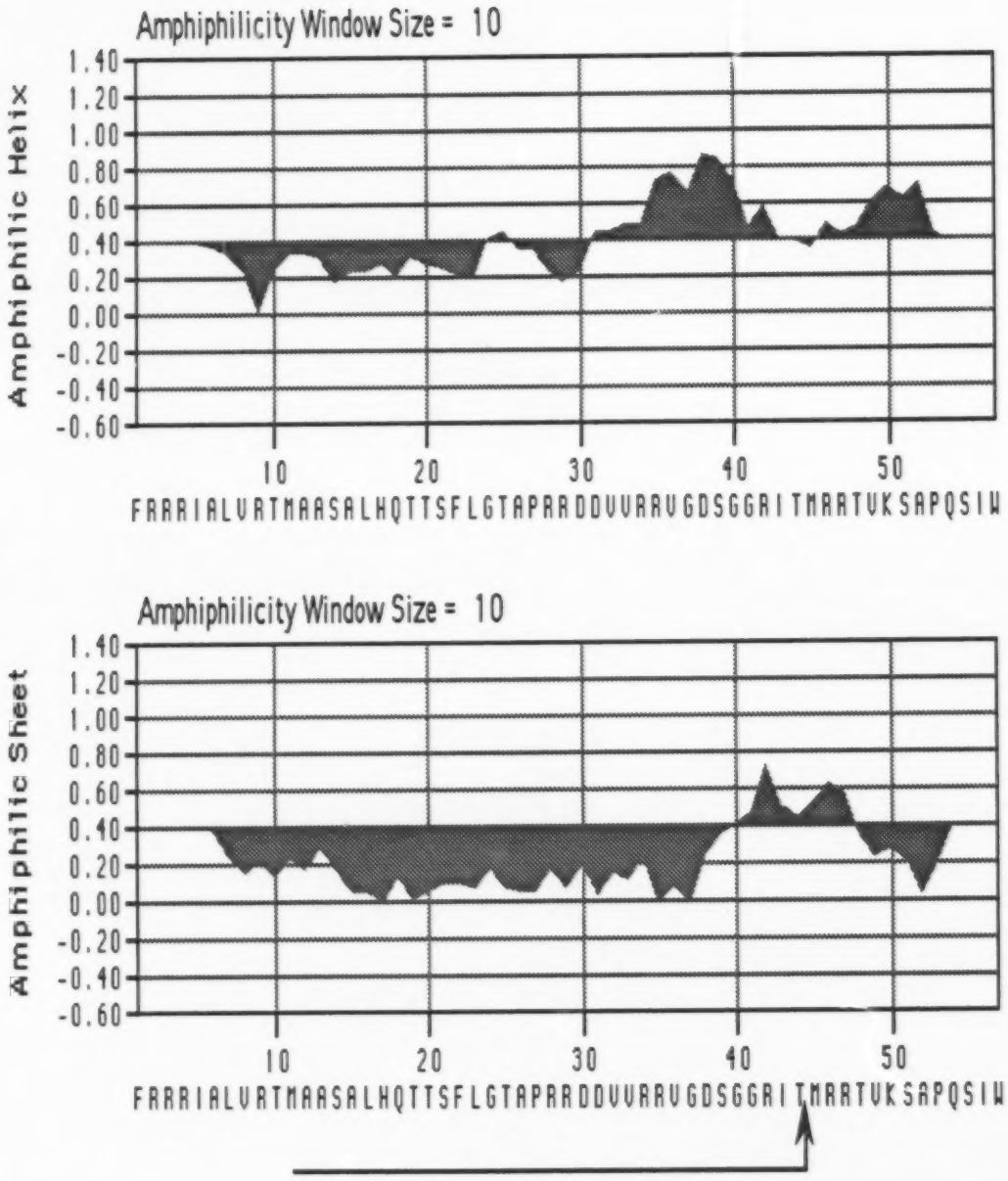


Figure 4.2C. The amphiphilic profile of the transit peptide for the rice2 *cab* gene. This is a Type II *cab* gene from PSII. The proposed transit peptide sequence is underlined and the arrow indicates the proposed cleavage site.

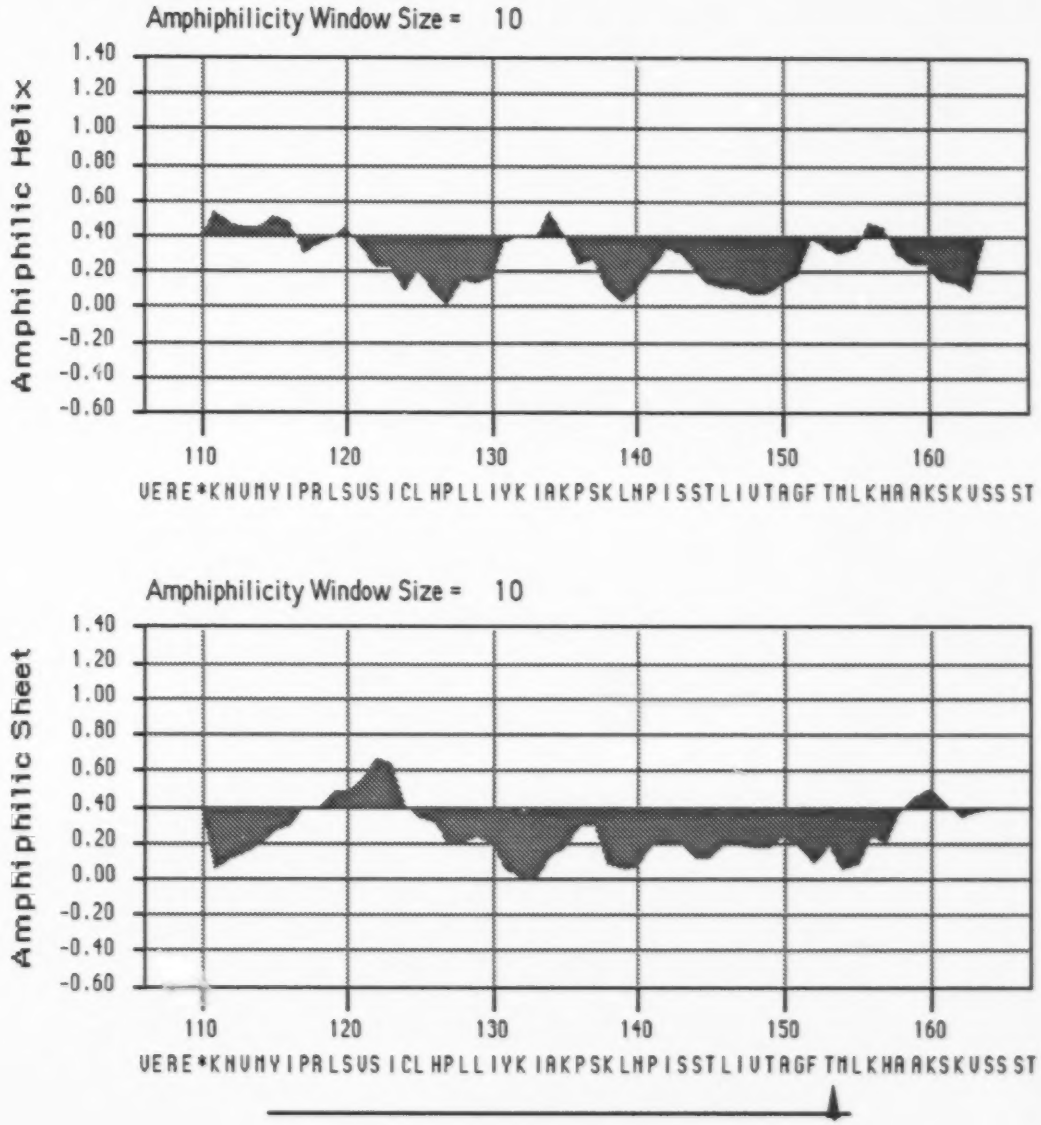


Figure 4.2D. Amphiphilic profile of an apple Type I *cab* gene. The transit peptide sequence is underlined and the proposed cleavage site is denoted with an arrow.

DISCUSSION

CAB Transit Peptides

As sequence evidence accumulates, the number of different CAB transit peptides also increases. Within PSII there now appear to be a least four groups of transit peptides based on protein sequence (Table 4.1).

Type I CAB Transit Peptides

The most common class includes those transit peptides that transport Type I CAB proteins across the chloroplast membrane. The peptide size is well conserved at 33 amino acids; however, lengths range from 32 residues (soybean3; Walling et al., 1988) to 36 residues (pea80; Cashmore 1984). The amino end of the peptide seems to be more conserved with strong regions of consensus at the first three positions and residues 6 through 10. Residue 4 seems to vary in 40% of the aligned individuals, while residue 5 changes in only two copies; however, the region surrounding it appears to include several insertions. The changes in positions 10 through to the cleavage site seem to involve species-specific substituitions, so that it is possible to group copies from the same species together based on their specific changes. For example, the last three amino acids before the cleavage site for all soybean (4) copies are Arg-Val-Ser; however, in *Arabidopsis* (3 copies) the same sequence is Arg-Val-Thr, in petunia (5 copies) this becomes Lys-Ala/Val-Thr and in tomato (6 copies) it is Arg-Ile-Thr.

It is not possible to distinguish each copy separately based on amino acid changes.

For example, soybean 4 and 5 have exactly the same transit peptide protein sequence. This is also the case with tomato copies 1B and 1C and *Arabidopsis* 165 and 180. And while there are no nucleotide changes between soybean 4 and 5, there are 5 silent third position substitutions between both tomato 1B and 1C, and *Arabidopsis* 165 and 180.

Type II CAB Transit Peptides

The second most common class are those transit peptides which transport the Type II CAB proteins into the chloroplast (Table 4.1). These include seven copies from six species and range in size from 31 amino acids (Lemna 19; Karlin-Neumann et al., 1985) to 37 residues (*Pinus thunbergii*; Yamamoto et al., 1988). All of these copies differ by at least three amino acid changes.

Woody Plant Transit peptides

The third group contains three copies from two pine species and one from a pear tree from what seem to be Type I copies of CAB. These peptides are much longer, 40 or 44 residues long, and have only two regions that seem to be similar to the other type I copies. The CAB copy from lodgepole pine differs by only one amino acid from one copy from scots pine (pINE13). These two species differ from the other scots pine copy by nine substitutions (Table 4.1). As well, pear and pINE9 have a similar insertion (position 9) and deletion (position 32). An interesting features within this group is the occurrence of four amino acids M-A-T-T preceeding the transit peptide of pINE9 and pear. It is not clear whether this functions in the amino terminal portion of the peptide.

In contrast, lodgepole pine and scots pine (pINE13) have T-A-A-T preceeding their amino terminal portion in the same location. This short sequence is similar to the amino terminal portion of the Type I copies which have M-A-A-A/S-T and nearly identical to the M-A-A-T found in maize (Matsouka et al., 1987) and petunia 22R (Dunsmuir 1985). It would seem that if this short sequence was important in a regulatory or structural sense then it would be under strong selection pressure.

Transit Peptides of Lower Plants

This final group is a collection of all the lower plants that do not seem to fit any of the groups already listed because of significant protein sequence differences. Within this group there are substantial differences, both in size and in sequence, of the transit peptides. The smallest is *Chlamydomonas reinhardtii* (Imbault et al., 1988) with a proposed transit peptide of 18 residues and increases to the algae *Dunaliella salina* (Long et al., 1987) with a proposed transit peptide is twice as large at 40 amino acids. Although the amino terminal portion of the two *Chlamydomonas* species (Imbault et al., 1988; Larouche et al., 1991) are very similar, they differ in both overall length and in the composition of the carboxy portion of the peptide (Table 4.1).

The transit peptide of the moss *Physcomitrella patens* (Long et al., 1987) is similar, both in length and amino acid sequence, to the functional transit peptide of the fern, *Polystichum munitum* (Pichersky et al., 1990). Clearly, the closest similarity resides in the last 18 residues of the carboxy portion of their respective transit peptides.

The two most unusual transit peptides belong to the algae, *Dunaliella salina* (Long

et al., 1989), and to the domestic apple, *Malus x domestica* (Chen et al., 1990). Both are 40 amino acids long and neither shares obvious similarity to each other or to any of the other groups. This may mean that these *cab* genes are incorrectly assigned to LHCII.

Comparison of The Transit Peptides Between the Groups

Visual comparison of the transit peptide sequences across the four groups is difficult and results in very few similarities. The most prominent feature seems to be the difference in size between the woody plants, with 40 to 44 amino acids and the other members of PSII, with approximately 33-35 residues. This range in size appears in the lower plants as well, as described above.

Computer alignment programs (i.e. MacVector) are unable to detect any significant sequence similarity between the pines and Type I or Type II transit peptides. Nevertheless, there do seem to be two short but well conserved regions in most of the transit peptides in each of the groups. The first region is a short hydrophobic sequence, F/L-A-G, that occurs in 33 of the 42 (79%) transit peptides listed in Table 4.1. A further 5 of the 42 copies vary slightly from this consensus motif. The F/L-A-G is replaced by F-V-G (occurs in two), F-A-F (occurs in one), L-V-G (occurs once) and F-S-G (once). If the variants are grouped with the main motif then 90% of PSII transit peptides contain this conserved region. Of the 4 species with no apparent homology, three of these are algae and may therefore be expected to be well diverged from higher plants. Nevertheless, both Chlamydomonids have a similar motif, F-A-L, although it is located much closer to the amino end of the peptide. Whether this is a significant similarity is

difficult to speculate, but it is appealing.

The most striking anomaly from this consensus motif is its lack of occurrence in apple. The overall peptide sequence diverges significantly from all other higher plants and this itself is surprising. There is a short hydrophobic sequence in apple (P-L-L-I) that occurs in approximately the same location as the consensus motif in the other transit peptides. This is the only hydrophobic sequence larger than two amino acids in the first 33 residues. It would seem that if this F-A-G consensus region is important, then in apple, the role would be filled by the sequence P-L-L-I.

The other conserved region occurs at the carboxy terminal end of the transit peptide. This region is the site of cleavage where the stromal peptidase acts to cleave the transit peptide from the mature protein. Therefore, it would seem likely that this region should be well conserved. Purified stromal peptidase for converting the prSSU to the mature SSU has been shown to be capable of processing several different precursors (Schmidt and Mishkind, 1986). Thirty nine (93%) of the transit peptides terminate with the consensus sequence (G/A)-(R/K)-X-(S/T) with the functional form of H-B-H-P amino acids. Where H,B,P represent hydrophobic, basic and polar amino acids respectively, as defined by Cantor and Schimmel (1980). As well, this particular arrangement of amino types does not occur at any other location within these transit peptides. It is possible that the stromal peptidase recognizes this particular arrangement and cleaves at the first such structure it comes across. This would suggest that this arrangement of amino acids, with specific properties, would be strongly selected.

There are three transit peptides that do not fit this scheme and they are algae

Chlamydomonas moewsii and *C. reinhardtii*, and the apple, *Malus x domestica*. The *Chlamydomonids* do not have this motif anywhere in their transit peptides and their proposed transit peptide terminate with B-B-H-H-H. Finally, the anomalous apple terminates with H-H-H-P. The conserved motif, H-B-H-P, does occur twice in apple but not near the carboxy terminal end. This motif occurs at the amino end, at positions 4 to 7 in the transit peptide and again in the middle at residues 21 to 24.

It has been suggested that chloroplast transit peptides generally contain the motif (A/I)-X-(A/C) at the cleavage processing site (Gavel and Von Heijne 1990). I have found no evidence for this specific amino acid sequence in the CAB transit peptides of PSII. Nor does this H-X-(H/P) motif agree with the observations presented above. Therefore this general chloroplast transit peptide does not fit any of the CAB transit peptides of PSII.

Mechanisms of Transport

There have been several reviews covering the most probable hypotheses dealing with protein transport and routing in chloroplasts (Smeekens et al., 1990, Keegstra et al., 1989, Keegstra 1989). The lack of amino acid similarity between chloroplast transit peptides, as outlined above, suggests two, possibly related interpretations of apoprotein transport into and within chloroplasts. First, the observation of sequence conservation within each type of transit peptide, and the apparent lack of homology between types, has prompted suggestions of multiple receptor sites located on the chloroplast surface and the interaction of the different transit peptides with different receptors (Keegstra 1989,

Keegstra et al., 1989). Conversely, it may be that the important information may not reside in the primary structure of the transit peptide. This latter possibility may involve some higher order structure (Keegstra 1989) or suggests that transit peptides are devoid of any regular secondary or tertiary structure. It has been suggested that they are essentially random coils interacting with molecular chaperones to pass into the chloroplast (Von Heijne and Nishikawa, 1991).

Model of Transportation For LHC

I propose that the transport of all PSII LHC into chloroplasts are receptor mediated and that the recognition sequence involves the short conserved central hydrophobic (F/L)-A-G motif. Apoproteins would bind to the chloroplast surface upon recognition of this motif. The mature protein would be transported through the chloroplast membrane without the complete transit peptide passing through the envelope. The large hydrophobic region in pine may serve as a transmembrane domain which holds the apoprotein in place until a stromal peptidase releases the mature peptide by cleaving the transit peptide away. At this point the mature peptide passes on to the thylakoid for insertion into that structure, while the transit peptide is degraded.

I believe that this could be tested by site directed mutagenesis or deletion analysis of the (F/L)-A-G motif.

In summary, the CAB transit peptides from PSII exhibit some of the general features common to other chloroplast transit peptides, such as a predicted amphiphilic beta-strand at the C-terminal end and a general deficiency of acidic amino acids.

Nevertheless, it does not appear that CAB transit peptides as a whole share the common cleavage motif, nor are they as rich in hydroxylated amino acids as predicted from other cTPs.

The CAB transit peptides exhibit a wide range of sizes and are difficult to align. I suggest that they fall into at least four groups which share two small regions of homology, a (F/L)-A-G motif and strongly conserved structural cleavage site.

I propose that lodgepole pine, and perhaps the other CAB transit peptides, bind to a chloroplast surface receptor, import the mature peptide into the stroma and are cleaved off without completely entering the chloroplast.

LITERATURE CITED

Batschauer A., B. Ehmann and E. Schaefer. 1991. Cloning and characterization of a chalone synthase gene from mustard and its light-dependent expression. Plant Mol. Biol. 16:175-185.

Boutry, M., F. Nagy, C. Poulsen, K. Aoyagi and N.-H. Chua. 1987. Targeting of bacterial chloramphenicol acetytransferase to mitochondria in transgenic plants. Nature 328:340-342.

Cantor, C.R., and P.R. Schimmel. 1980. *Biophysical Chemistry. Part I: The Conformation of Biological Macromolecules*. W.H. Freeman and Co., San Francisco.

Cashmore, A. 1984. Structure and expression of a pea nuclear gene encoding a chlorophyll a/b-binding polypeptide. Proc. Nat. Acad. Sci. 81:2960-2964.

Castresana, C., R. Staneloni, V.S. Malik and A.R. Cashmore. 1987. Molecular characterization of two clusters of genes encoding the Type I CAB polypeptides of PSII in *Nicotiana plumbaginifolia*. Plant Mol. Biol. 10:117- 126.

Chen, H., S.S. Korban, and D.E. Buetow. 1990. Nucleotide sequence of an apple nuclear gene encoding a light-harvesting chlorophylla/b binding polypeptide of photosystem II. Nucl. Acids Res. 18:679.

Chua, N.-H., and G.W. Schmidt. 1979. Transport of proteins into mitochondria and chloroplasts. J. Cell Biol. 81:461-483.

Demmin, D.S., E.J. Stockinger, Y.C. Chang and L.L. Walling. 1989. Phylogenetic relationships between the chlorophyll a/b binding protein (CAB) multigene family: an intra- and interspecies study. J. Mol. Evol. 29:266-279.

Dunsmuir, P. 1985. The petunia chlorophyll a/b binding protein genes: a comparison of *Cab* genes from different gene families. Nucl. Acids Res. 13:2503-2518.

Eisenberg, D., R.M. Weiss and T.C. Terwilliger. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Nat. Acad. Sci. 81:140-144.

Gavel, Y. and G. Von Heijne. 1990. A conserved cleavage-site motif in chloroplast transit peptides. FEBS Letters 261:455-458.

Greenland, A.J., M.V. Thomas and R.M. Walden. 1987. Expression of two nuclear genes encoding chloroplast proteins during early development of cucumber seedlings. Planta 170:99-110.

Imbault, P., C. Wittemer, U. Johanningmeier, J.D. Jacobs and S.T. Howell. 1988. Structure of the *Chlamydomonas reinhardtii cab*II-1 gene encoding a chlorophyll a/b-binding protein. Gene 73:397-407.

Jansson, S., and P. Gustafsson. 1990. Type I and Type II genes for the chlorophyll a/b-binding protein in the gymnosperm *Pinus sylvestris* (Scots pine):cDNA cloning and sequence analysis. Plant Mol. Biol. 14:287-296.

Karlin-Neumann, G.A., B.D. Kohorn, J.P. Thornber and E.M. Tobin. 1985. A chlorophyll a/b-protein encoded by a gene containing an intron with characteristics of a transposable element. J. Mol. Appl. Genet. 3:45-61.

Karlin-Neumann, G.A. and E.M. Tobin. 1986. Transit peptides of nuclear-encoded chloroplast proteins share a common amino acid framework. EMBO J. 5:9-13.

Keegstra, K. 1989. Transport and routing of proteins into chloroplasts. Cell 56:247-253.

Keegstra, K., L.J. Olsen and S.M. Theg. 1989. chloroplastic precursors and their transport across the envelope membranes. Ann. Rev. Plant Physiol. Plant Mol. Biol. 40:471-501.

Klee, H.J., Y.M. Muskopf and C.S. Gasser. 1987. Cloning of an *Arabidopsis thaliana* gene encoding 5-enolpyruvylshikimate-3-phosphate synthesis: sequence analysis

and manipulation to obtain glyphosated tolerant plants. Mol. Gen. Genet. 210:437-442.

Kohorn, B., E. Harel P.R. Chitnis, J.P. Thornber and E.M. Tobin. 1986. Functional and mutational analysis of the light-harvesting chlorophyll a/b protein of thylakoid membranes. J. Cell Biol. 102:972-981.

Kyte, J., and R.F. Doolittle. 1982. A simple method for displaying the hydrophathic character of a protein. J. Mol. Biol. 157:105-132.

Lamppa, G.K. 1988. The chlorophyll a/b binding protein inserts into the thylakoids independent of its cognate transit peptide. J. Biol. Chem. 263:14996-14999.

Lamppa, G.K., G. Morelli and N-H Chua. 1985. Sturcture and developmental regulation of a wheat gene encoding the major chlorophyll a/b-binding polypeptide. Mol. Cell. Biol. 5:1370-1378.

Larouche, L., C. Tremblay, C.Simard, and G. Bellemare. 1991. Characterization of a cDNA encoding a PSII-associated chlorophyll a/b-binding protein (CAB) from *Chlamydomonas moewusii* fitting into neither type I nor type II. Curr. Genet. 19:285-288.

Leutwiler, L.S., E.M. Meyerowitz and E.M. Tobin. 1986. Structure and expression of three light-harvesting chlorophyll a/b-binding protein genes in *Arabidopsis thaliana*. Nucl. Acids Res. 14:4051-4064.

Long, Z., S. Wang and N. Nelson. 1989. Cloning and nucleotide sequence analysis of genes coding for the major chlorophyll-binding protein of the moss *Physcomitrella patens* and the halotolerant alga *Dunaliella salina*. Gene 76:299-312.

Mason, J.G. 1989. Nucleotide sequence of a cDNA encoding the light-harvesting chlorophyll a/b binding protein from spinach. Nucl. Acids Res. 17:5387.

Matsuoka, M. 1990. Classification and characterization of cDNA that encodes the light-harvesting chlorophyll a/b binding protein of photosystem II from rice. Plant Cell Physiol. 31:519-526.

Matsuoka, M., Y. Kano-Murakami and N. Yamamoto. 1987. Nucleotide sequence of cDNA encoding the light-harvesting chlorophyll a/b binding protein from maize. Nucl. Acids Res. 15:6302.

Pichersky, E., R. Bernatzky, S.D. Tanksley, R.W. Breidenbach, A.P. Kausch and A.R. Cashmore. 1985. Molecular characterization and genetic mapping of two clusters of genes encoding chlorophyll a/b-binding proteins in *Lycopersicon*

esculentum (tomato). Gene 40:247-258.

Pichersky, E., N.E. Hoffman, V.S. Malik, R. Bernatzky, S.D. Tanksley, L. Szabo and A.R. Cashmore. 1987. The tomato *Cab*-4 and *Cab*-5 genes encode a second type of CAB polypeptides localized in Photosystem II. Plant Mol. Biol. 9:109-120.

Pichersky, E., D. Soltis and P. Soltis. 1990. Defective chlorophyll a/b-binding protein genes in the genome of a homosporous fern. Proc. Nat. Acad. Sci. 87:195-199.

Schmidt, G.W., and M.L. Mishkind. 1986. The transport of proteins into chloroplasts. Ann. Rev. Biochem. 55:879-912.

Smeekens, S., J. Van Oosten, M. de Groot and P. Weisbeek. 1986. Silene cDNA clones for a divergent chlorophyll a/b-binding protein and a small subunit of ribulose bisphosphate carboxylase. Plant Mol. Biol. 7:433-440.

Smeekens, S., P. Weisbeek and C. Robinson. 1990. Protein transport into and within chloroplasts. TIBS 15:73-76.

Stayton, M.M., M. Black, J. Bedbrook and P. Dunsmuir. 1986. A novel chlorophyll a/b binding (*Cab*) protein gene from petunia which encodes the lower weight *Cab* precursor protein. Nucl. Acids Res. 14:9781-9796.

Von Heijne, G. and K. Nishikawa. 1991. Chloroplast transit peptides: the perfect random coil? FEBS Lett. 278:1-3.

Von Heijne, G. and C. Manoil. 1990. Membrane proteins: from sequence to structure. Prot. Engineer. 4:109-112.

Von Heijne, G., J. Steppuhn and R.G. Herrmann. 1989. Domain structure of mitochondrial and chloroplast targeting peptides. Eur. J. Biochem. 180:535-545.

Walling, L.L., Y.C. Chang, D.S. Demmin and F.M. Holzer. 1988. Isolation, characterization and evolutionary relatedness of three members from the soybean multigene family encoding chlorophyll a/b binding proteins. Nucl. Acids Res. 16:10477-10492.

Yamamoto, N., M. Matsuoka, Y. Kano-Murakami, Y. Tanaka and Y. Ohashi. 1988. Nucleotide sequence of a full length cDNA clone of light harvesting chlorophyll a/b binding protein gene from green dark-grown pine (*Pinus thunbergii*) seedling. Nucl. Acids Res. 16:11829.

CHAPTER FIVE

PHYLOGENETIC ANALYSIS OF *Cab* GENES

INTRODUCTION

The nuclear genes of the *cab* family code for an apoprotein comprised of a transit peptide and mature polypeptide which are transported into the chloroplast. The precursor is cleaved off and the mature protein is routed and inserted into the thylakoid membrane. The chlorophyll a/b binding protein (CAB) is involved in the formation of a light harvesting protein complex (LHC-II) which involves three identical CAB polypeptides, eight chlorophyll *a* molecules, seven chlorophyll *b* molecules and several carotenoids (Butler and Kuhlbrandt, 1988). Together these light harvesting complexes bind 50% of the total chlorophyll on earth (Kuhlbrandt and Wang, 1991). Light energy is collected by the LHC-II and transferred into the photosynthetic reaction centres I and II.

The *cab* gene family is interesting because it has been shown to include different gene types encoding different polypeptides for each of the two photosystems (PSI and PSII). Within photosystem II (PSII), the most common family comprises the Type I genes which encode higher molecular weight proteins than Type II.

The occurrence of several types of CAB proteins in each of the photosystems emphasizes the diversity of the *cab* gene family. Best elucidated in tomato, the *cab* family now includes three *cab* Types in PSII, four Types in PSI, as well as the related core antenna, CP24 and Cp29 in PSII (Green et al., 1991). Several groups have demonstrated that there are shared amino acid blocks conserved between these LHC (Green et al., 1991).

An earlier phylogenetic study involving 25 *cab* gene sequences from seven dicotyledonous species and two monocotyledonous species found that maximum likelihood analysis supported established morphological phylogenies (Demmin et al., 1989). They concluded that the Solanaceae were more distantly related to monocots than were the Brassicaceae and Fabaceae. As well they suggested that their data demonstrated that formation of the Type I and II *cab* genes occurred before the divergence of the dicots and monocots.

This study involves the reexamination of the phylogeny of the PSII *cab* genes by adding several new dicotyledonous and monocotyledonous sequences, as well as including *cab* gene sequences from conifers and lower plants. The large data set has been analyzed using cladistic and phenetic methods. The results indicate that the Type I and Type II cab genes arose before the angiosperm/gymnosperm split approximately 300 million years ago (Raven et al., 1981). A pre-Type I and II *cab* gene, derived from an ancestral gene, may have been similar to the sequence which presently resides in lower plants like moss and fern.

MATERIALS AND METHODS

PAUP Analysis

Forty PSII *cab* gene sequences encoding the mature CAB protein were compiled from the literature and Genbank searches. All sequences were aligned using MacVector 2.0. Only the 1st and 2nd codon position changes were used for parsimony analysis.

These sites should be selectively constrained and therefore any changes should be more informative than third position changes which are usually neutral. The lodgepole pine Type I *cab* gene PNcAB2 was used as the standard against which all other published *cab* genes were compared. There were 124 informative sites out of possible 470 1st and 2nd codon nucleotides. Informative sites were bases which differed, at that particular nucleotide, in at least two other *cab* copies. The complete data set (i.e. 40 sequences) was tested using PAUP 3.0g (Swofford, 1991).

The region of greatest variability among the *cab* genes occurs within the first 55-65 nucleotides of the 5' end. Within this region, alignment is difficult and often requires visual adjustment and assignment of gaps. In an effort to avoid researcher biased alignments, a second analysis was performed on the informative sites covering the 640 nucleotides outside this variable region. This reduced the number of informative sites located at the 1st or 2nd codon positions to 101. Both data sets were analyzed using a general heuristic search in which the taxa were added randomly and the branches were swapped by the tree bisection-reconnection procedure (TBR). This method is the most effective of the branch swapping methods in obtaining the shortest tree (Swofford, 1991). Tree bisection and reconnection involves the bisection of a tree along a branch yielding two disjointed subtrees. The subtrees are then reconnected by rejoining the subtrees at a different location. The result is a new tree. This procedure is repeated until all possible bisections and pairwise reconnections are evaluated (Swofford, 1991). The analysis was replicated five times; the shortest trees are saved and a consensus tree is produced.

Distance Matrix

Sequence pairwise distance estimates were calculated for 20 sequences, using DNADIST from PHYLIP Ver3.3 (Felsenstein, 1990), for 20 sequences. This program uses Kimura's 2-parameter model (Kimura, 1980) in which the base frequencies are assumed to be equal and the transition/transversion ratio is set at 2. Distances are a measure of sequence divergence and are scaled in terms proportion of expected number of nucleotide substitutions as a function of time (Felsenstein, 1990).

Both Type I and II sequences were included, and where possible comparison was made between types from the same species covering representative dicotyledonous, monocotyledonous and coniferous members. Several lower plant sequences were also analysed.

Distance matrices were analzsed using FITCH and KITSCH (PHYLIP Ver. 3.3) to produce trees with informative branch lengths. FITCH is a program that carries out Fitch-Margoliash and Least Squares calculations for distance matrices and estimates branch length. KITSCH is the same as FITCH, however it assumes a molecular clock, produces a rooted tree and estimates branch length and time of divergence (Felsenstein, 1990).

Model Assumptions

For the DNA parsimony models it is assumed (Felsenstein 1990) that:

1. each site evolves independently,

2. different lineages evolve independently,

3. the probability of a base substitution at a given site is small over the lengths of time involved in a branch of the phylogeny, and

4. expected amounts of change in different branches of the phylogeny do not vary.

For the distance matrix programs based on distances estimated from molecular data it is implicitly assumed that:

1. each distance is measured independently from the others; no item of data contributes to more than one distance,

2. distance between each pair of taxa is drawn from a distribution with an expectation which is the sum of values (in effect amount of evolution) along the tree from one tip to the other. The variance of the distribution is proportional to a power *p* of the expectation.

RESULTS

DNA Parsimony

DNA parsimony infers the ancestral sequence between two extant taxa and by choosing the minimum number of mutational changes to describe the observed tree topology (Nei 1987). Forty sequences (Table 5.1) were analysed using PAUP and compared at 124 informative sites. *Physcomitrella patens* (moss) was used as an outgroup to root the tree. A cladogram representing the evolutionary relatedness of the *cab* genes was produced by a general heuristic search where the sequences were added randomly until the shortest tree was found. All the trees produced from this search are then compared and a single consensus tree is produced (Fig. 5.1). This represents

Table 5.1. Taxonomic classification[1] of the plants used in this study. The species are divided into their Divisions, followed by their Classes, Families and species names.

DIVISION	CLASS	FAMILY	SPECIES	CLONE	(REF)[2]
SPERMATOPHYTA					
ANGIOSPERMAE					
Liliopsida (monocots)	Arales	Lemnaceae	*Lemna gibba*	Lemna 19	(1)
				Lemna 30	(2)
	Cyperales	Poaceae	*Triticum*	Wheat 1.6	(3)
			Hordeum	Barley	(*)
			Zea mays	Maize	(4)
			Oryza	Rice1,Rice2	(5)
Magnoliospida (dicots)	Solanales	Solanaceae	*Lycopersicon esculentum*	Tomato 1B, 3C, 4, 5	(6,7)
			Petunia hybrida	Petunia 3,13,22R,22L,25 91, 37	(8-10)
			Nicotiana plumbaginifolia	Tobacco A,B	(11)
	Fabales	Fabaceae	*Glycine max*	Soybean 2-4, CPA	(12,13)
			Pisum sativm	Pea 80	(14)
	Rosales	Rosaceae	*Pyrus pyrofolia*	Pear	(*)
			Malus x domestica	Apple	(15)
	Violales	Cucurbitaceae	*Cucumis sativus*	Cucumber A,B	(16)
	Capparales	Brassicaceae	*Arabidopsis thaliana*	Arabidopsis 2,3	(17)
			Sinapis alba	Mustard	(18)
	Caryophyllales	Caryophyllaceae	*Silene pratensis*	White Campion	(19)
GYMNOSPERMAE	Coniferae	Pinaceae	*Pinus contorta*	Lodgepole Pine, PNcAB2	this study
			Pinus sylvestris	Scots Pine pINE9, pINE11, pINE13	(20)
			Pinus thunbergii	Japanese Pine, PINETHUN	(21)
PTERIDOPHYTA	Filicales	Polypodiaceae	*Polystichum munitum*	Fern3	(22)
BRYOPHYTA			*Physcomitrella patens*	Moss	(23)
Others			*Chlamydomonas reinhardtii*		(24)
			C. moewusii		(25)
			Dunaliella salina		(23)

1. Based on Cronquist, 1981.

2. References: (1) Karlin-Neumann et al., 1985; (2) Kohorn et al., 1986; (3) Lamppa et al., 1985; (4) Matsuoka et al., 1987; (5) Matsuoka, 1990; (6) Pichersky et al., 1985; (7) Pichersky et al., 1987; (8) Dunsmuir et al., 1983; (9) Dunsmuir 1985; (10) Stayton et al., 1986; (11) Castresana et al., 1987; (12) Walling et al., 1988; (13) Demmin et al., 1989; (14) Cashmore, 1984; (15) Chen et al., 1990; (16) Greenland et al., 1987; (17) Leutwiler et al., 1986; (18) Batschauer et al., 1991; (19) Smeekens et al., 1986; (20) Jansson and Gustafsson, 1990; (21) Yamamoto et al., 1988; (22) Pichersky et al., (1990); (23) Long et al., 1989; (24) Imbault et al., 1988; (25) Larouche et al., 1991. * indicates unpublished sequence pulled from EMBL Sequence Library.

the one most parsimonious tree found after five replications. Several interesting features are immediately apparent. First, all the Type II sequences containing dicot, monocot, conifer and lower plant species form a distinct cluster separate from all the proposed Type I sequences. Second, all the Type I sequences from dicotyledonous species cluster in a large central group clearly separate from the Type I sequences from the monocotyledonous species and the conifers. Third, it is clear that all Type II sequences from different species are more closely related to each other than they are to the Type I copies from the same species. This trend occurs in the dicots, monocots and pines that are represented. Fourth, several ambiguities exist among the clusters. For example, pear (*Pyrus pyrofolia*) and apple (*Malus x domestica*), both in the Family Rosaceae, should be very closely related; however, it appears that, based on nucleotides sequences, they are significantly different. Pear is apparently more closely related to pines than it is to any of the angiosperms. This association brings into question the accuracy of the pear sequence. In contrast, apple seems to be most similar to the cluster containing the crucifer species. There is also an unusual association of the fern, *Polystichum munitum*, joining the cluster containing the Type I *Brassica* species sequences. This is strange because I believe that the fern *cab* copy is a Type II sequence. Therefore it would be expected to cluster either within the larger Type II cluster, or because of its evolutionary age, below this clade. Nevertheless, it appears to cluster with the angiosperm Type I *cab* sequences, suggesting that it is more closely related to Type I *cab* genes than to Type II genes.

Removal of the 23 informative sites from the highly variable 5' end of the cab

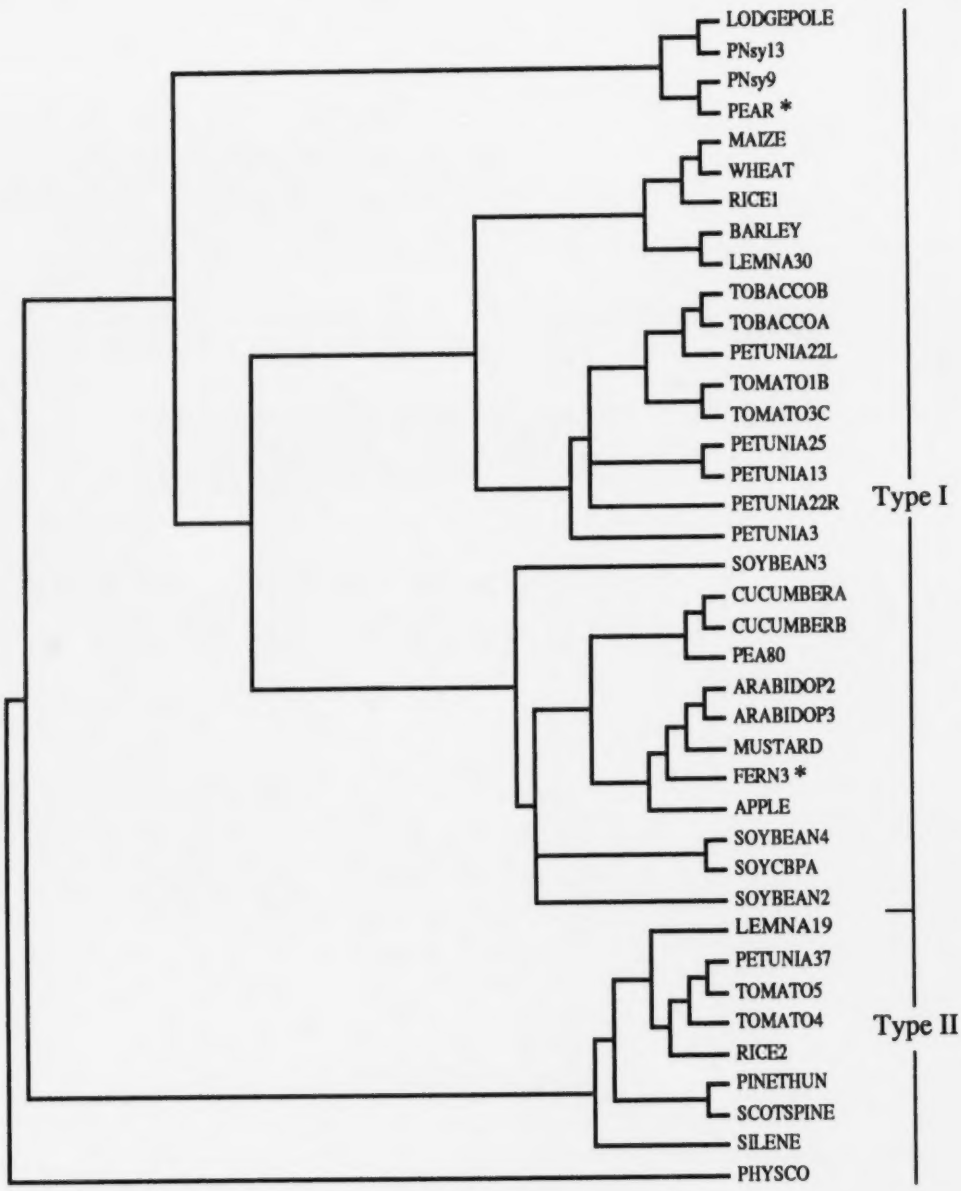


Figure 5.1. Rectangular cladogram representing the phylogenetic relationships of the PSII *cab* genes. This is a consensus tree produced from a general heuristic search based on 124 informative sites after five replications and a shortest tree length of 580. *Physcomitrella patens* was used as the outgroup. The different *cab* gene types are indicated. * denotes the Type II Fern sequence and the unusual pear sequence.

gene results in 101 informative sites in 40 sequences. These were analyzed in exactly the same method as above with *Physcomitrella patens* as the outgroup (Fig. 5.2). These results are very similar to the analysis performed on the larger data set. The most significant difference is the removal and splitting of the Type I monocots from the larger central cluster found in Fig.5.1. The consequence of this is the formation of several branches each terminating in monocot species (Fig. 5.2). This cladogram presents a more appealing biological explanation of the development of the *cab* genes. Based on site differences the dicots split from the monocot line containing wheat. As well, the monocots arose from a line that led to the present day pines. A second difference is that the apple sequence clusters with the pea, the soybean and cucumber copies. Both pea and soybean are members of the Fabaceae family. These species, of the species represented in this study are the most closely related, taxonomically, to the family Rosaceae to which apple is a member. Both families belong to the Subclass Rosidae (Cronquist, 1981). The Type II gene branching is the same in both analyses. Finally, the moss and fern form a cluster that branches off the conifer line. When the three algae species, *Chlamydomonas reinhardtii*, *C. moewusii*, and *Dunaliella salina*, are included there are only two structural differences in the topology of the tree in Figure 5.1. Again the apple sequence clusters with the pea, the soybean and cucumber copies. Both pea and soybean are members of the Fabaceae family. Second, the algae form a small cluster branching off the Type II pine sequence (Fig.5.3). This alignment suggests that algae and pine shared a common ancestor more recently than did the conifers and angiosperms. This is difficult to accept because of the supposed evolutionary age of the algae. Any alignment

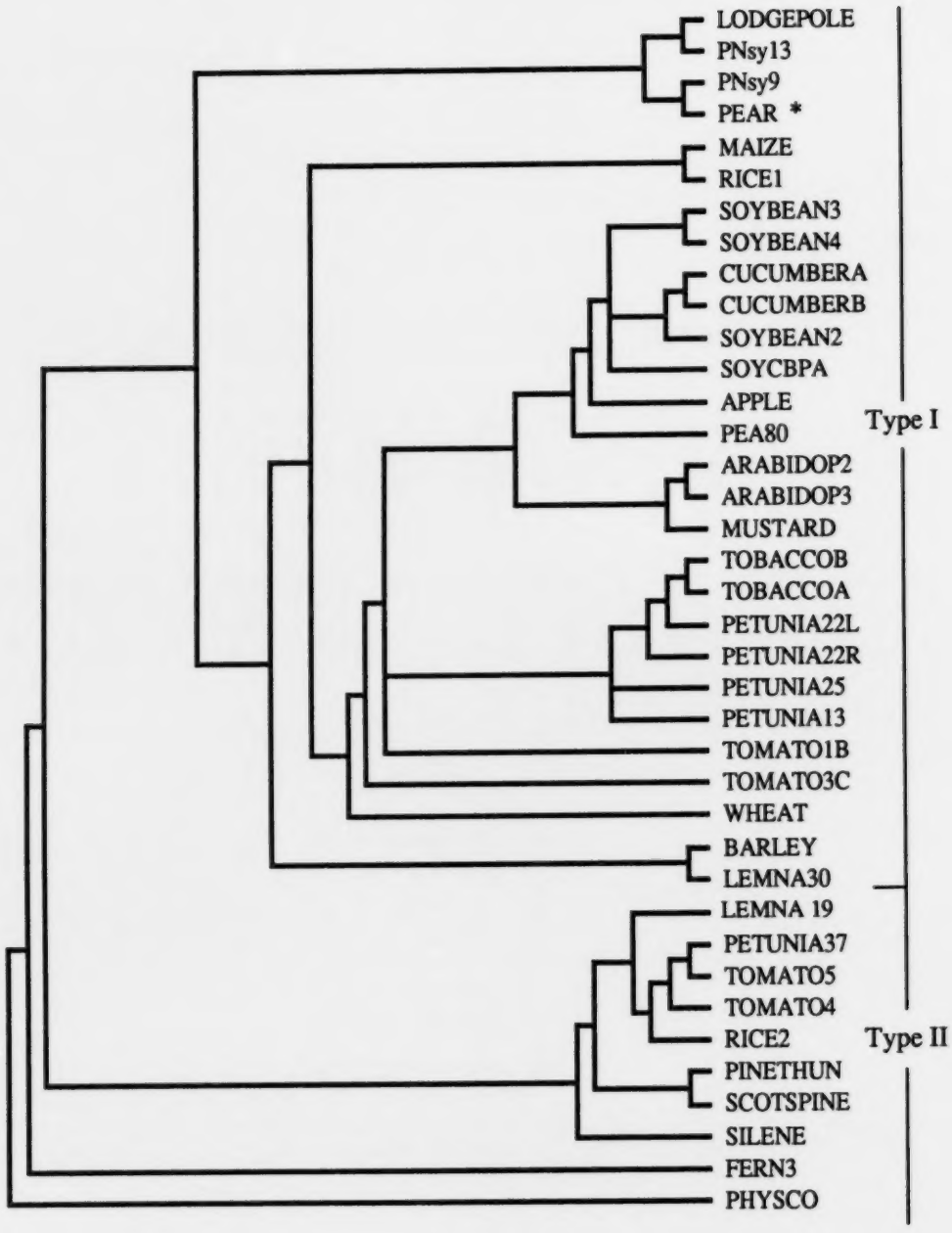


Figure 5.2. Phylogenetic relationships of the PSII *cab* genes. DNA parsimony analysis based on 101 informative sites. This the consensus tree after five replications of a general heuristic search with the random addition of samples. The tree is rooted with the moss *Physcometrella patens* and has a length = 453. The * denotes the unusual pear sequence.

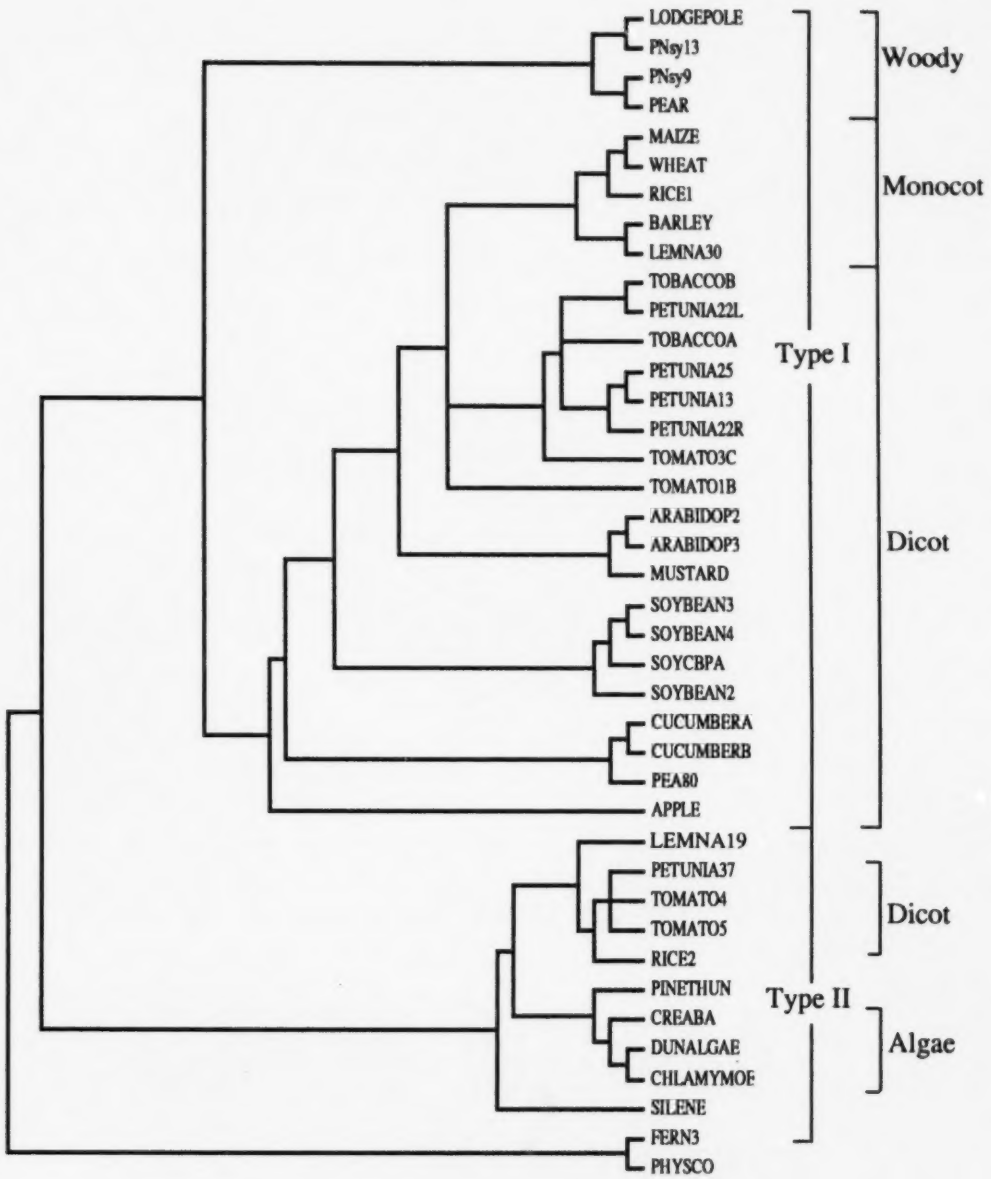


Figure 5.3. Consensus tree constructed using 124 informative sites after five replications and the random addition of taxa. The algae species were included and the *Physcomitrella patens* was used as the outgroup.

based on evolutionary complexity would likely place the algae on a separate branch, while the ferns and their allies would be more closely aligned with the seed plants.

Distance Analysis

In contrast to parsimony methods, phenetic analysis aligns taxa based on phenotypic, or in the case of molecular sequence data, genotypic similarities. In this study pairs of sequences were compared and an estimate of the genotypic similarity is calculated. The observed frequency of substitutions is then estimated or measured based on these distance estimates. For a discussion of this approach see Nei (1987). The observed frequency of substitutions is then compared to the expected frequency and taxa are clustered based on the observed deviation from expectation. In other words, the greater the observed substitutions from the expected value, the farther the two taxa are apart.

Analysis of twenty cab sequences representing lower plants, conifers and angiosperms using the Fitch-Margoliash Model suggests several trends which differ from parsimony analysis. First, contrary to parsimony, the major clusters in this analysis are comprised of all dicot sequences aligning in one cluster and all other taxa aligning in a second major cluster. Second, within each cluster the Type I sequences form clusters separately from the Type II sequences (Fig. 5.4). Third, the moss and fern, both Type II copies, join the cluster containing angiosperms, while the algae species are in the same

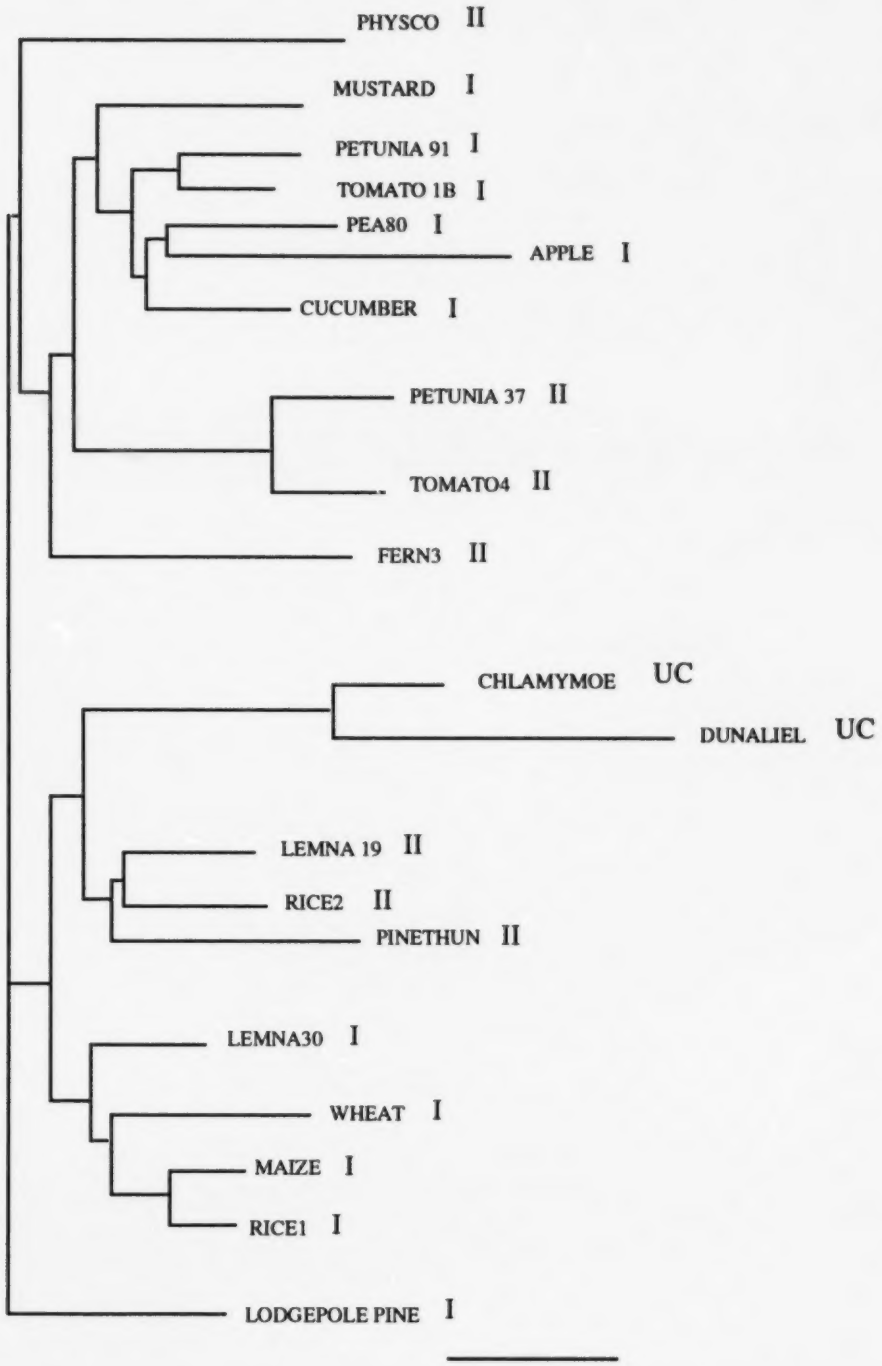


Figure 5.4. Phenetic analysis of 20 selected *cab* gene sequences. This gene tree is derived from the FITCH program in PHYLIP Ver.3.3 and tests pairwise distance data using the Fitch-Margoliash model. 714 nucleotide positions were examined. The bar represents 10% divergence for unit of time. UC stands for uncertain designation, I for Type I cab gene and II for Type II gene.

cluster as are the conifers and monocots.

If the data is reanalyzed assuming a molecular clock and therefore all species are the same distance from the root then the results are only slightly different (Fig. 5.5). Instead of two major clusters containing dicots in one cluster and, monocots and conifers in the other, there is only a single cluster containing the Type I dicots and the Type I and II monocots and conifers. The Type II copies from dicots represent a small secondary splinter cluster. The fern and moss sequences secure these clusters and the algae *Dunaliellia salina* roots the tree. What is most surprising is the appearance of apple close to the root of the tree placed between two algal species. A strict interpretation of this suggests that the ancestral line giving rise to apple separated from algae and maintained a single evolutionary path separate from other angiosperms. This would be biologically awkward and leads to further suspicion of the apple sequence.

DISCUSSION

There are several approaches to constructing phylogenetic trees. Of these, the two most common are the phenetic and cladistic methods. The phenetic or distance method constructs trees based on the phenotypic similarity of the species involved. This method is merely a grouping of organisms and does not consider the evolutionary patterns of the species. Nevertheless, the phenetic approach is justified when the rate of evolutionary change is not well known or when interested in the taxonomy based on phenotypic resemblance only.

The second approach, the cladistic method includes the study of the pathways of

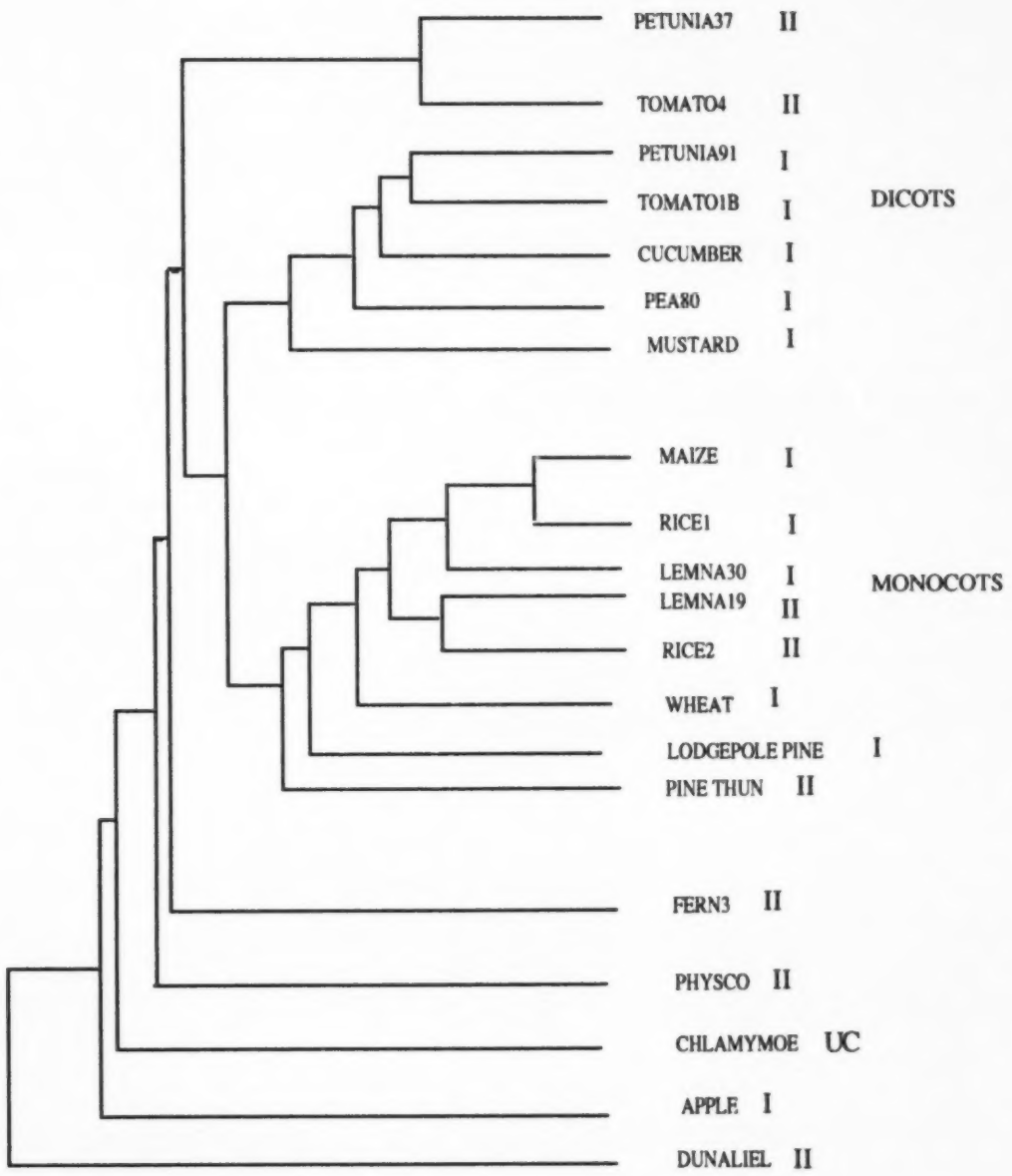


Figure 5.5. Phylogenetic tree derived from KITSCH program of PHYLIP Ver.3.3. KITSCH assumes a molecular clock. Analysis performed on 20 selected *cab* gene sequences examined at 714 nucleotide positions.

evolution. Trees are constructed by considering various possible ways of evolution and following certain rules and choosing the best possible tree. For example, the number of branches among a group of organisms, which branch connects to which other branch and what is the branching sequence are characteristics determined by this approach. This is not necessarily a better approach because the model's assumptions are not always satisfied with molecular data. For example, multi-gene families offer an example of a lineage that may not evolve independently. It has been shown that members of multigene families tend to retain more similarity than would be expected. They evolve together and this homogenizing process is known as concerted evolution (Hartl and Clark 1989). In concerted evolution the copies of a gene family in one species may be more similar to each other than to the same copies in another species. This is because of the genetic interactions of unequal crossing-over and/or gene conversion result in a multigene family evolving as a unit in a concerted fashion (Arnheim 1983).

The question of the evolution of the *cab* genes is an interesting problem. Cladistic analysis of Type I and II *cab* sequences from a wide range of plant types suggest that the split between the two PSII *cab* Types occurred before the split between the angiosperms and gymnosperms. The maintenance of separate clades, composed exclusively of Type I or Type II sequences, supports this hypothesis. Within each clade the *cab* Types are represented by conifer, monocotyledonous and dicotyledonous species. Clearly Type I *cab* genes are more similar between such distantly related species as lodgepole pine and tomato than are Type I and Type II *cab* sequences within the same species. This phenomenon is observed in the monocots *Lemna* and rice, the dicots petunia and tomato

and in pine, a conifer. The suggestion of an early appearance of the split between Type I and II *cab* genes in the angiosperms before the dicot/monocot split, has been suggested before (Demmin et al., 1989). Nevertheless, their analysis was based on only two monocotyledonous species, *Lemna* and wheat. In this study I have included three new monocot Type I sequences and one additional Type II gene from a monocot. As well, I have incorporated several pine species. These include Type I *cab* genes from lodgepole and scots pine and a Type II *cab* gene from *Pinus thunbergii* and *P. sylvestris*. New Type I *cab* genes from several dicotyledonous species such as cucumber, mustard and apple have been included to increase the scope of earlier analysis. Of further interest is the addition of *cab* gene sequences of lower plants. It is intriguing that fern and moss, which both appear to have introns in the exact location as other PSII Type II *cab* genes, contain a mix of the diagnostic amino acids (Demmin et al., 1989; Jansson and Gustafsson, 1990) that enable separation of Type I and Type II *cab* copies. For example, of the fourteen "Type specific" amino acids (see Table 1.2, pg. 10), moss contains eight Type I specific residues and six Type II amino acids and fern has eight Type I, three Type II and three other amino acids. This suggests at least two possibilities. Either this amino acid diagnostic system for separating PSII *cab* Types is only specific for gymnosperms and angiosperms and/or that plants below the gymnosperm/angiosperm split contain an ancestral *cab* gene(s) which contain features of both the future Type I and Type II *cab* genes including introns and a pre-split mix of diagnostic amino acids. This may help explain how a Type II lower plant would be clustered with the dicotyledonous Type I *cab* genes instead of more closely related species such as moss or even conifers.

Phenetic analysis of the same sequences generally supports the observations from cladistic analysis. Of interest is the arrangement of the major clusters after assuming a molecular clock (i.e. KITSCH in PHYLIP Ver. 3.3). In contrast to parsimony analysis in which all Type I sequences above the gymnosperm/angiosperm split clustered together, phenetic analysis produces major clusters in which all dicots aligned together and all conifers and monocots group together. This is to be expected because phenetic analyses are based on associating phenotypic similarities together without regard to evolutionary pathways (Nei, 1987). It seems possible that at the nucleotide level, conifers are more similar to monocots than to dicots, thereby resulting in the clustering of conifers and monocots. Within each cluster there is still the observed separation of the Type I and Type II sequences, however the clusters now appear to resolve into a phylogenetic tree more representative of the one presented in an earlier study (Demmin et al., 1989).

In conclusion I have constructed gene trees based on the two most common phylogenetic methods. In both analyses there were clear distinctions between the Type I and Type II *cab* genes. Parsimony analysis clustered all species irrespective of evolutionary distance together into clades that were separated by the *cab* gene type. Phenetic analysis resulted in dicot species aligning together and monocots and conifers grouping together. Within each cluster the Type I genes were separated from the Type II genes. These results suggest that the Type I/Type II split evolved before the separation of the gymnosperms and angiosperms. As well, further evidence suggests that fern and moss maintain ancestral *cab* genes that contain features of both Type I and Type II cab genes.

Future projects should concentrate on filling in specific gaps in the proposed tree. This might include isolating and sequencing *cab* genes from other conifers such as *Abies* and *Picea*, and other gymnosperms. Several lycopods and bryophytes would also be informative. Finally, the two most unusual sequences are from two fruit trees from the Family Rosaceae, therefore information about other members of Rosaceae would be useful as would sequences from several other woody angiosperms.

LITERATURE CITED

Arnheim, N. 1983. Concerted evolution of multi-gene families. pp38-61. *In* M. Nei and R.K. Koehn (eds) *Evolution of Genes and Proteins*. Sinauer Associates, Sunderland, MA.

Batschauer A., B. Ehmann and E. Schaefer. 1991. Cloning and characterization of a chalone synthase gene from mustard and its light-dependent expression. Plant Mol. Biol. 16:175-185.

Butler, P.J.G. and W. Kuhlbrandt. 1988. Determination of the aggregate size in detergent solution of the light-harvesting chlorophyll a/b-protein complex from chloroplast membranes. Proc. Nat. Acad. Sci. 85:3797-3801.

Cashmore, A. 1984. Structure andexpression of a pea nuclear gene encoding a chlorophyll a/b-binding polypeptide. Proc. Nat. Acad. Sci. 81:2960-2964.

Castresana, C., R. Staneloni, V.S. Malik and A.R. Cashmore. 1987. Molecular characterization of two clusters of genes encoding the Type I CAB polypeptides of PSII in *Nicotiana plumbaginifolia*. Plant Mol. Biol. 10:117- 126.

Chen, H., S.S. Korban, and D.E. Buetow. 1990. Nucleotide sequence of an apple nuclear gene encoding a light-harvesting chlorophylla/b binding polypeptide of photosystem II. Nucl. Acids Res. 18:679.

Cronquist, A. 1981. *An Integrated System of Classification of Flowering Plants.* Columbia University Press, New York.

Demmin, D.S., E.J. Stockinger, Y.C. Chang and L.L. Walling. 1989. Phylogenetic relationships between the chlorophyll a/b binding protein (CAB) multigene family: an intra- and interspecies study. J. Mol. Evol. 29:266-279.

Dunsmuir, P. 1985. The petunia chlorophyll a/b binding protein genes: a comparison of Cab genes from different gene families. Nucl. Acids Res. 13:2503-2518

Dunsmuir, P., S.M. Smith and J. Bedbrook. 1983. The major chlorophyll a/b binding protein of petunia is composed of several polypeptides encoded by a number of distinct nuclear genes. J. Mol. Genet. 2:285-300.

Felsenstein, J. 1990. PHYLIP Manual Version 3.3. University Herbarium, University of California, Berkeley, California.

Green, B.R., E. Pichersky and K. Kloppstech. 1991. Chlorophyll a/b-binding proteins: an extended family. Trends In Biochem. Sci. 16:181-186.

Greenland, A.J., M.V. Thomas and R.M. Walden. 1987. Expression of two nuclear genes encoding chloroplast proteins during early development of cucumber seedlings. Planta 170:99-110.

Imbault, P., C. Wittemer, U. Johanningmeier, J.D. Jacobs and S.T. Howell. 1988. Structure of the *Chlamydomonas reinhardtii cab*II-1 gene encoding a chlorophyll a/b-binding protein. Gene 73:397-407.

Jansson, S., and P. Gustafsson. 1990. Type I and Type II genes for the chlorophyll a/b-binding protein in the gymnosperm *Pinus sylvestris* (Scots pine):cDNA cloning and sequence analysis. Plant Mol. Biol. 14:287-296.

Karlin-Neumann, G.A., B.D. Kohorn, J.P. Thornber and E.M. Tobin. 1985. A chlorophyll a/b-protein encoded by a gene containing an intron with characteristics of a transposable element. J. Mol. Appl. Genet. 3:45-61.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120.

Kohorn, B., E. Harel P.R. Chitnis, J.P. Thornber and E.M. Tobin. 1986. Functional and mutational analysis of the light-harvesting chlorophyll a/b protein of thylakoid membranes. J. Cell Biol. 102:972-981.

Kuhlbrandt, W., and D.N. Wang. 1991. Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. Nature 350:130-134.

Lamppa, G.K., G. Morelli and N-H Chua. 1985. Sturcture and developmental regulation of a wheat gene encoding the major chlorophyll a/b-binding polypeptide. Mol. Cell. Biol. 5:1370-1378.

Larouche, L., C. Tremblay, C.Simard, and G. Bellemare. 1991. Characterization of a cDNA encoding a PSII-associated chlorophyll a/b-binding protein (CAB) from *Chlamydomonas moewusii* fitting into neither type I nor type II. Curr. Genet. 19:285-288.

Leutwiler, L.S., E.M. Meyerowitz and E.M. Tobin. 1986. Structure and expression of three light-harvesting chlorophyll a/b-binding protein genes in *Arabidopsis thaliana*. Nucl. Acids Res. 14:4051-4064.

Long, Z., S. Wang and N. Nelson. 1989. Cloning and nucleotide sequence analysis of genes coding for the major chlorophyll-binding protein of the moss *Physcomitrella patens* and the halotolerant alga *Dunaliella salina*. Gene 76:299-312.

Li, W-H and D. Graur. 1991. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Matsuoka, M. 1990. Classification and characterization of cDNA that encodes the light-harvesting chlorophyll a/b binding protein of photosystem II from rice. Plant Cell Physiol. 31:519-526.

Matsuoka, M., Y. Kano-Murakami and N. Yamamoto. 1987. Nucleotide sequence of cDNA encoding the light-harvesting chlorophyll a/b binding protein from maize. Nucl. Acids Res. 15:6302.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Pichersky, E., R. Bertanzky, S.D. Tanksley, R.W. Breidenbach, A.P. Kausch and A.R. Cashmore. 1985. Molecular characterization and genetic mapping of two clusters of genes encoding chlorophyll a/b-binding proteins in *Lycopersicon esculentum* (tomato). Gene 40:247-258.

Pichersky, E., N.E. Hoffman, V.S. Malik, R. Bernatzky, S.D. Tanksley, L. Szabo and A.R. Cashmore. 1987. The tomato *Cab*-4 and *Cab*-5 genes encode a second type of CAB polypeptides localized in Photosystem II. Plant Mol. Biol. 9:109-120.

Pichersky, E., D. Soltis and P. Soltis. 1990. Defective chlorophyll a/b-binding protein genes in the genome of a homosporous fern. Proc. Nat. Acad. Sci. 87:195-199.

Raven, P.H., R. Evert and H. Curtis. 1981. *Biology of Plants*. 3rd Edition. Worth Publishers, Inc. New York, New York.

Smeekens, S., J van Oosten, M. de Groot and P Weisbeek. 1986. Silene cDNA clones for a devergent chlorophyll a/b-binding protein and a small subunit of ribulosebiophosphate carboxylase. Plant Mol. Biol. 7:433-440.

Stayton, M.M., M. Black, J. Bedbrook and P. Dunsmuir. 1986. A novel chlorophyll a/b binding (*Cab*) protein gene from petunia which encodes the lower weight *Cab* precursor protein. Nucl. Acids Res. 14:9781-9796.

Swofford, D.L. 1991. PAUP:phylogenetic analysis using parsimony, Version 3.0. Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

Walling, L.L., Y.C. Chang, D.S. Demmin and F.M. Holzer. 1988. Isolation, characterization and evolutionary relatedness of three members from the soybean multigene family encoding chlorophyll a/b binding proteins. Nucl. Acids Res. 16:10477-10492.

Yamamoto, N., M. Matsuoka, Y. Kano-Murakami, Y. Tanaka and Y. Ohashi. 1988. Nucleotide sequence of a full length cDNA clone of light harvesting chlorophyll a/b binding protein gene from green dark-grown pine (*Pinus thunbergii*) seedling. Nucl. Acids Res. 16:11829.

CHAPTER SIX

GENERAL DISCUSSION AND CONCLUSIONS

The small multigene family which encodes the chlorophyll a/b binding protein (CAB) has many features for interesting and important studies. The genes encode proteins that bind up most of the chlorophyll into a complex involved in light energy capture and transfer to the Photosystems. This gene family is clearly important in photosynthesis; however, more remarkable is the complexity of its organization. Within each photosystem, there are several types of *cab* genes. For example, in tomato, researches have found three different types in PSII and four types in PSI. As well, it has been shown that other antenna molecules, such as CP24 and CP29 are related and that they, and the recently discovered early light inducible proteins (ELIPs), share a common amino acid framework that revolves around conserved internal blocks that probably arose through gene duplication (Green et al., 1991).

All the sequenced *cab* genes contain introns except for the Type I genes from PSII. The number and position of the introns vary among the different *cab* genes Types. Researchers have found 11 intron positions among the other cab genes and only three positions are common in more than one type of *cab* gene. In general, *cab* genes are significantly diverged (60-70%). This, however is not the case for the Type I and II genes of PSII which are only 10-15% diverged.

Another interesting aspect of the *cab* genes is that they are nuclear genes which are translated in the cytoplasm and then the CAB protein is transported into the chloroplast where it is shunted to the thylakoid membranes. The polypeptide binds up chlorophyll molecules, captures light energy and tranfers it to the photosystems.

All *cab* genes encode for a transit peptide which assists the mature peptide in crossing the chloroplast membrane. Available amino acid sequence data suggests that these transit peptides are well conserved between species for a specific type of *cab* gene. Nevertheless, there is little homology of the transit peptides between the types within the same species.

The evolution of the *cab* genes, particularly those of PSII, is a fascinating riddle. Phylogenetic analysis has suggested that the appearance of the Type I and II genes occurred before the split between the dicots and monocots.

These are some of the stimulating features of the *cab* genes which prompt exciting questions for research. I set out to examine the *cab* genes in the conifer lodgepole pine. Much work had been done on angiosperms, however there was no information on the *cab* genes on gymnosperms or lower plants. The availability of a genomic library of lodgepole pine made this species a logical choice. The objectives of this study were as follows:

1. To construct a genomic library from an individual of lodgepole pine and to screen with a homologous probe for *cab* genes.

2. To identify and characterize a homologous *cab* gene(s) from lodgepole pine and to contrast it to other published *cab* genes.

3. To compare and contrast the transit peptide deduced amino acid sequences from all known PSII *cab* genes with the hope of identifying any common features.

4. To phylogenetically analyze the nucleotide sequence of the coding region of all known *cab* PSII genes. These objectives have been met and outlined in the preceeding chapters.

To reiterate, a single copy of a *cab* gene was isolated from a genomic library from lodgepole pine using a genomic clone from *Lemna gibba* (LgAB19). This uncharacterized lodgepole pine *cab* clone was used as a homologous probe to screen a genomic library constructed from a single individual of lodgepole pine. A single *cab* copy was detected and sequenced. Based on the gene structure, the lack of an intron and sequence similarity to other *cab* genes it was determined that this was a Type I *cab* gene. The nucleotide sequence was highly similar to other Type I genes, however it was most closely related to another Type I *cab* gene from scots pine. Sequence similarity was maintained outside the coding region into the 5', and to a lesser extent, the 3' untranslated regions. Comparison of the promoter region from lodgepole pine to regulatory regions in other Type I *cab* genes produced few regions of similarity and no extensive homology in the important regulatory elements. This is, perhaps, not surprising since comparison of the regulatory regions from wheat do not conform to similar elements in pea or tobacco. As well, pea and tobacco do not contain significant regions of homology in their respective promoter regions. This lack of promoter specificity for the *cab* genes means that promoter deletion analysis will be necessary to identify the regulatory elements. This will be both important and demanding work.

Alignment of the transit peptides of the PSII *cab* genes was attempted with a

possible explanation of the groupings. As well, two conserved motifs were identified, one of possible importance in chloroplast membrane receptor recognition and the second a general consensus scheme for stromal peptidase recognition and cleavage.

Phylogenetic analysis of the published PSII *cab* gene sequences produced evidence of a *cab* gene duplication and the resultant formation of Type I and II genes that may have occurred before the gymnosperm/angiosperm split. In addition, some evidence is presented to consider the interesting placement of fern and moss in the phylogenetic trees.

Future Research

The question concerning the size of the *cab* gene family in lodgepole pine has not been resolved. It seems clear, based on angiosperm species, that there are generally more than three copies of the Type I *cab* gene in PSII and one to two copies of the Type II genes. There is insufficient information on the other *cab* gene types to forecast their occurrence in lodgepole pine. Perhaps a larger genomic library would permit resolution of the other copies.

Other coniferous species need to be screened for the identification of their *cab* genes. Are the *cab* genes so well conserved in conifers as indicated between lodgepole and scots pine? As well, it may be that only one *cab* gene for each gene type has evolved in conifers. This may be the direct result of their evergreen state. In a deciduous conifer species like *Larix laricina* the number of *cab* genes may be larger to reflect the need for large quantaties of the CAB protein because of each new year's needle production and initiation of photosynthesis. In conifers which retain their needles through

out the winter, production of CAB may start as soon as it is physiologically required. Therefore not as many copies of each gene would be required. Nevertheless, it will be necessary to test this hypothesis by screening both the respective genomes and testing the translation levels of different *cab* genes, during different seasons and at different times.

Characterization of the promoter region is a logical and important next step in the description of the *cab* genes in pines. Based on promoter analysis in other conifers it might be expected that the pine *cab* promoter regions may be similar to those analyzed in wheat. It has been shown that conifers recognize monocot but not dicot promoters (Good et al., 1990).

The developmental expression of *cab* genes has not been approached in conifers. Nevertheless, if it could be demonstrated that the *cab* genes belong to a multigene family with several members, it would be possible to determine if the different copies were specific for a developmental stage or tissue type.

Analysis of the transit peptides would be an important study for understanding the recognition signals and transport across the chloroplast membrane. Deletion analysis of the transit peptide sequence, specifically the F-A-G region, and fusion of mutant apoproteins into expression vectors would allow monitoring of the affect of short deletions on the efficiency and validity of transfer.

Finally it seems important to locate and study the internal recognition sequence in the mature protein which has been proposed as directing the protein into the thylakoid membranes. If identified, the next question is whether it is membrane specific or do all *cab* genes have the same structural sequence for their particular membranes.

LITERATURE CITED

Good, A.G., F. Bekkaoui, G. Pilante, D.I. Dunstan and W.L. Crosby. 1990. Anaerobic induction in conifers: expression of endogenous and chimeric anaerobically-induced genes. Physiolog. Plantarum 78:441-446.

Green, B.R., E. Pichersky and K. Kloppstech. 1991. Chlorophyll a/b-binding proteins: an extended family. Trends In Biochem. Sci. 16:181-186.